A Capacity Allocation Planning Model for Integrated Care and Access Management

Jivan Deglise-Hawkinson[1] • **Jonathan E. Helm**[2] • **Todd Huschka**[3] • **David L. Kaufman**[4] • **Mark P. Van Oyen**[5*]

1 jivan@umich.edu; Revenue Management – Operations Research, American Airlines, Fort Worth, TX

2 helmj@indiana.edu; Operations & Decision Technologies, Indiana University, Bloomington, IN

3 huschka.todd@mayo.edu; Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, MN

4 davidlk@umich.edu; Management Studies, University of Michigan–Dearborn, Dearborn, MI

5 vanoyen@umich.edu; Industrial & Operations Engineering, University of Michigan, Ann Arbor, MI

[*] Corresponding author: vanoyen@umich.edu

Abstract: The prevailing first-come-first-served approach to outpatient appointment scheduling ignores differing urgency levels, leading to unnecessarily long waits for urgent patients. In data from a partner healthcare organization, we found in some departments that urgent patients were inadvertently waiting longer for an appointment than non-urgent patients. This paper develops a capacity allocation optimization methodology that reserves appointment slots based on urgency in a complicated, integrated care environment where multiple specialties serve multiple types of patients. This optimization reallocates network capacity to limit access delays (indirect waiting times) for initial and downstream appointments differentiated by urgency. We formulate this problem as a queueing network optimization and approximate it via deterministic linear optimization to simultaneously smooth workloads and guarantee access delay targets. In a case study of our industry partner we demonstrate the ability to (1) reduce urgent patient mean access delay by 27% with only a 7% increase in mean access delay for non-urgent patients, and (2) increase throughput by 31% with the same service levels and overtime.

Keywords: Advance Planning and Scheduling; Access Waiting Time Management; Health Care Operations; Patient Flow Modeling; Mixed Integer Programming; Queueing Networks.

History: Submitted: March 2016; R1: May 2017; R2: May 2018; Accepted: July 2018

1 Introduction

Patient health and financial concerns have spurred a growing shift to delivery of outpatient care through coordinated care networks (American Hospital Association 2015). There is also growing interest in the ability to limit patients' wait times to receive an appointment. Still, there is not much literature on methods to limit access delays that are set according to patient type, and the problem is compounded for networks of outpatient specialist services. We develop methods that are general in the numbers of services and patient types. A case study of three departments at our industry partner is used to demonstrate how to balance the access delays for (1) new patients who have an unknown condition that requires rapid diagnosis through a stochastic series of consults in multiple medical specialties, and (2) established patients who are involved in ongoing monitoring and treatment of a previously diagnosed condition. We call the patients who present with a new condition *urgent* patients. Aside from organizational priorities that may justify urgent status, new patients require a diagnosis and therefore rapid access to mitigate their health risks. Urgency in the outpatient setting differs from urgent care in the inpatient setting where most critical patients are admitted through the emergency department.

Coordinated care networks are faced with the challenge of providing rapid access to urgent patients. To do so, they reserve some of their capacity for initial diagnostic visits. However, reserving capacity for new patients in one department, if not carefully considered, can lead to long delays for established patients returning for a follow-up appointment (we call these patients *non-urgent*) in that department and possibly for other patients (urgent and non-urgent) in other departments. Our definition of urgency comes from our partner organization, however it can be easily tailored to other definitions based on the needs of the application. This capacity management and allocation problem is especially challenging due to the interconnected network of services employed in a diagnostic *itinerary* and the complex relationships between the delays for different patient types. Next we discuss the main contributions of the paper in terms of application, theory, and management insights.

Application: This model addresses the challenges of care networks as they seek to stratify patient access delay (also called indirect waiting in Gupta and Denton (2008)) according to the needs of each patient type. In our case study of a partner organization, we show that it is possible to improve mean access delay for urgent patients while limiting delays for non-urgent patients. We extend this analysis, to show that the distribution on access delay is also controllable to fit the needs of the organization by allowing for multiple service level constraints. Next, we show how our model can be used to create Pareto curves that illustrate the tradeoffs between three key competing metrics: throughput, overtime, and access delay. Finally, we demonstrate the value of the integrated solution, showing that the siloed approach can cause significant downstream congestion which can be mitigated by our integrated model.

Technical: The technical contributions of this paper include novel methods for capacity planning and allocation across an integrated network of care services. Specifically, we formulate and solve a capacity reservation optimization through the analysis and linearization of a complex (and non-traditional) queueing network that accounts for multiple patient classes, multiple specialties, and multiple competing metrics. Our performance metrics described in Sec. 3, if captured exactly, are nonlinear in the decision variables, so we transform a nonlinear stochastic queueing network into a tractable, deterministic linear optimization model. Finally, we develop methods for controlling not only the mean, but the full shape of the access delay distributions, which in turn shapes the workload distributions.

Managerial Insights: Without an integrated model, ad hoc or siloed approaches in one service often lead to unintended consequences for other services in the system. For example, increasing throughput in one department at the partner institution led to increased access delay and congestion in other departments. Even with clear goals in mind for competing metrics, powerful analytical methods and decision support are needed to tie these metrics to capacity planning decisions. The what-if scenario capabilities provided by our model can support a wide array of managerial decision frameworks.

The rest of this paper is organized as follows. Sec. 2 describes the problem context based on our collaboration with a leading integrated care provider that serves patients from throughout the United States and around the world. We use real data to demonstrate challenges and inefficiencies in the planning approach currently used in practice, and we review the literature. Sec. 3 describes the system dynamics and presents the model. Sec. 4 discusses the conversion of nonlinear, stochastic system dynamics into a set of deterministic, linear optimization equations. In Sec. 5, we numerically validate our model and present a case study of our partner organization. Sec. 6 concludes the paper.

2 Context and Literature Survey

Our focus is on outpatient capacity allocation planning models for integrated care. Integrated care has been identified as an increasingly important trend in the U.S. healthcare system (see Kocher and Sahni 2010). In contrast to scheduling, we perform planning through the optimization of an appointment *template*. Our partner institution, like many others, requests that medical departments reserve some slots for specific types of patients in their appointment template. Each slot may have a deterministic duration that depends on the type of appointment it is designated for. This template process has historically been managed in a siloed and reactive manner, whereas our method designs optimized templates so that managerial decisions become precise, integrative, and proactive.

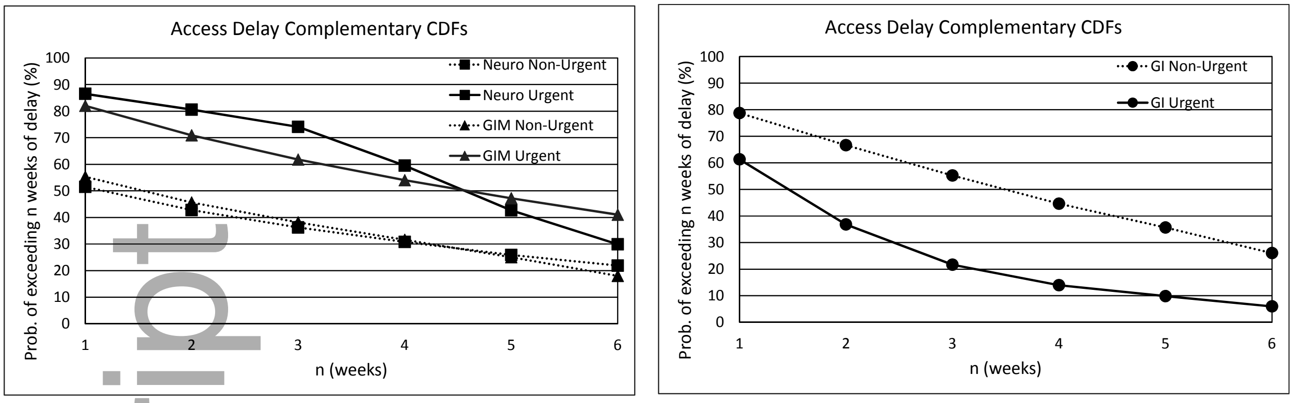
To illustrate our context, consider the General Internal Medicine (GIM) department. Patients requiring diagnosis and treatment planning for a new condition are typically scheduled for an initial/*root* appointment via the GIM template. Based on the analysis of this initial consult at GIM, *downstream* appointments are generated in GIM and other departments — e.g., Gastroenterology (GI) — for further analysis, diagnosis, and treatment. These downstream appointments are not known in advance of the initial consult, so some capacity in the template must also be reserved to accommodate each patient's dynamically generated itinerary. This is similar to the inpatient context in which a patient's treatment path is unknown at the time of arrival to the hospital (see Bekker and Koeleman 2011). However, in the business model for care networks such as the one we study, an itinerary typically consists of a rapid succession of multiple visits to multiple specialties over the course of several days. Hence, the GIM template reserves some capacity for root appointments that are scheduled in advance, and reserves the remaining capacity for two types of unplanned downstream visits: those originating from a root appointment in GIM – called *follow-up appointments*, and those originating from root appointments in other departments that refer their patients to GIM – called *internal referrals*.

In this context, a decision support system is needed for a variety of reasons. Suppose, for example, that hospital management anticipates an increase in the volume of urgent patients to the GIM department, and they were considering hiring new physicians to accommodate the new demand. There are several natural questions that might arise: How many physicians are needed? If the number of root appointments for urgent patients in GIM increases, how many downstream visits might be generated? That is, what is the distribution of the stochastic resource requirements during a patient's itinerary, which occurs from the the time of the root appointment until the patient leaves the care network? Moreover, how will increasing the number of patients in GIM affect other departments? Do increases in GIM cause unacceptable delays in, say, GI?

Based on roughly one year of data from our industry partner, Fig. 1 shows the historical complementary c.d.f. of access delay by patient urgency to obtain a root appointment in GI, GIM, and Neurology under the historical capacity plan. *Access delay* is the time between when a patient requests an appointment and when the appointment is able to be scheduled. Essentially, the graph in Fig. 1 shows the probability of exceeding n weeks of access delay ($n = 1, .., 6$). The percent of urgent vs. non-urgent patients in each department was: GI 25.9% urgent, GIM 71.6% urgent, Neuro 76.7% urgent.

Fig. 2 shows (1) the average daily resource capacity for each department (in physician hours), (2) the average daily total workload (in hours) based on current practice scheduling, and (3) the average daily workload in that department that is generated from internal referrals. The whiskers represent ± one standard deviation.

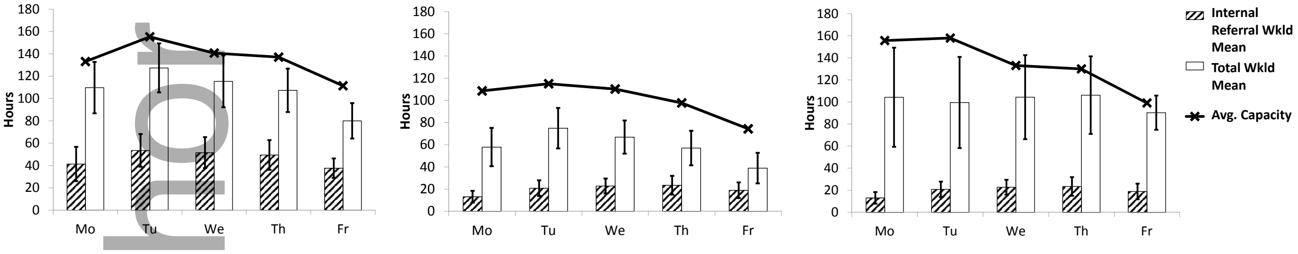
From Fig. 1, observe that urgent vs. non-urgent access delays vary significantly by medical department. This is not surprising since effective control of wait times for advance appointments is not well understood. One consequence is the surprising finding that urgent patients in both Neurology and GIM have stochastically longer delays to obtain a root appointment than non-urgent patients, which is seen in Fig. 1(a) where the complementary c.d.f.'s of urgent patients are strictly greater than those of non-urgent patients. Two-sample t-tests resulted in p-values of 0.000, indicating that the mean access delays are different for urgent and non-urgent patients.



(a) Neurology and GIM (b) GI

Figure 1: Historical complementary c.d.f.'s of access delay (in weeks) for the root appointment of urgent and non-urgent patients in (a) Neurology and GIM and (b) GI. For instance, for GIM, 54% of urgent patients wait 4 weeks or more, while only 32% of non-urgent patients wait 4 weeks or more. In contrast, for GI, while 45% of non-urgent patients wait 4 weeks or more, only 14% of urgent patients wait 4 weeks.

In contrast to Neurology and GIM, urgent GI patients (Fig. 1(b)) experience stochastically shorter delays than non-urgent patients. This is because the GI department recently began an initiative to prioritize urgent patients. However, this heuristic prioritization scheme resulted in much higher mean access delays for non-urgent patients (compared to GIM and Neurology). This initial effort, though, indicates an opportunity to optimize access delays with greater precision while working within the existing appointment scheduling framework of our partner institution, which allows patient slots to be restricted to certain types of patients.



(a) GIM (b) GI (c) Neurology

Figure 2: Current practice internal referral workload in physician hours and total workload in three medical departments. The bar heights are the historical workload means. The whiskers display ± one standard deviation.

Figure 2 demonstrates two key features that motivate our capacity planning methodology. First, internal referrals make up a significant proportion of the total workload for GIM. This workload cannot be controlled by GIM themselves, but can have a major impact on the access delay and overtime at GIM. This motivates the need for an integrated network capacity plan, since the internal referrals are indirectly controlled by root appointments in other departments. Second, workload is not particularly well matched to capacity. For example, Neurology is very likely to exceed capacity later in the week, while there is more slack capacity on Monday and Tuesday. This motivates the need for an optimization to smooth the workload relative to capacity across the days of the week, reducing access delay and overtime.

Our solution methodology is designed to account for the metrics of access delay, overtime, and utilization by controlling multiple demand streams across a network of services. Section 5.2 presents cases that include (1) the reducing mean access delay, (2) controlling the shape of the distribution of access delay by patient type, and (3) increasing the number of urgent cases served while meeting service levels for non-urgent patients and limiting overtime caused by downstream demands for subsequent visits in an interconnected network of departments. The last point is particularly new to the literature.

Next, we survey the most relevant literature. Emergency care has developed priority-based reactive admission control methods based on severity/urgency scoring during triage to reactively differentiate access delays based by patient severity; e.g., see Saghafian et al. (2014). In contrast, proactive advance planning methods for appointment-based service operations are lacking. To the best of our knowledge, the objectives of this research are beyond the capabilities of published research or available commercial products. Our approach differs significantly from the appointment-based scheduling literature (and other areas such as capacity planning, lead time quoting, and revenue management) because it contains multiple technically difficult features including: (1) scope (network vs. single clinic), (2) planning horizon (multi-day vs. single day), and (3) stochastic service itineraries in a network (as opposed to a single appointment).

Much of the outpatient scheduling and planning literature focuses on a single resource/clinic, often modeled as a queueing system, and considers scheduling patients to time slots within a day considering no-shows, doctor availability, etc., as in the key survey papers of Cayirli and Veral (2003), Gupta and Denton (2008), and Hulshof et al. (2012). Denton and Gupta (2008) were among the first to identify indirect wait, which we call access delay, as an important yet overlooked operational metric that negatively impacts patient outcomes and can be managed through appropriate planning and scheduling. The authors state that, in contrast to effective operational management methods, the "soft nature of provider capacity is relied upon to absorb variations in demand." Further, they point out that this problem inherently has a multi-day horizon without a clear decomposition approach. The practical value of capacity planning considering access delay is indicated in Vermeulen et al. (2009), which reports an operational implementation of a capacity reservation approach for a single resource (CT scanners). That paper focuses on the percentage of patients meeting their access delay target, which is achieved by dynamically adjusting the capacity with final adjustments performed by a human scheduler.

Aligning with most outpatient healthcare practice, our paper takes a multi-day planning approach, which is similar to "advance scheduling," in which patients are booked/scheduled into future days at the time of their arrival. Gerchak et al. (1996) provides an early stochastic dynamic programming analysis of a time-homogeneous surgical planning system that must optimize the amount of daily capacity to be reserved for emergency (same day) surgeries. Gupta et al. (2007) addresses elective surgery booking control and maximum access delay by patient class using a Markov decision process (MDP). In the context of operating rooms, Lamiri et al. (2008) emphasizes planning of elective cases known in advance under uncertain demand for emergency cases. Feldman et al. (2014) develops a heuristic for a daily appointment booking system, and utilizes a multinomial logit model to incorporate patient preferences.

Several recent studies for single-unit (non-network) systems have considered priority scheduling and dynamic capacity allocation problems solved via approximate dynamic programming (ADP) (e.g., Herbots et al. 2010). Some papers consider multi-priority jobs arriving dynamically that must be scheduled on some future date (or rejected) with holding costs for delays or overtime (see Patrick et al. 2008, Erdelyi and Topaloglu 2010, Gocgun and Puterman 2014). Patrick et al. (2008) uses ADP and heuristics to address a single diagnostic resource with stationary capacity (a perhaps simpler model than our template based approach), but does not explicitly incorporate access delay targets, or a care network with feedback. Their advance scheduling method incorporates elements of both real-time scheduling and planning. Patients requesting service can be scheduled immediately, diverted/rejected, or deferred with a later call-back to schedule the appointment. Their model has been simplified in some ways but also extended to n demand classes, random service times, and multiple resources in Truong (2015), which links advance scheduling to allocation scheduling. While diversions or deferment can be appropriate in some settings, most outpatient care requires an up-front appointment date and time. Gocgun and Puterman (2014) consider chemotherapy scheduling and also use ADP for an MDP model considering diversion and scheduling costs (without overtime limits, which we model). They decompose the problem using a two-stage process by which they promise an appointment date in advance, then at a later date specify the time on that day. Their paper involves both planning and real-time decision making, but focuses on scheduling only follow-up appointments, while our work must integrate the resource allocation for both new and follow-up visits. Our mixed integer programming (MIP) based approach benefits from the relatively easy incorporation of many constraints. Like their paper, ours also makes the case that the time-of-day details of scheduling can be resolved well enough that the daily level decisions of who will be seen on that day can be made well in advance.

Other papers consider the fact that each patient (job) may initiate a time series of appointments over multiple days with deadlines/time windows for downstream appointments (see Gocgun and Ghate 2012, Sauré et al. 2012, Hulshof et al. 2013). Targeting clinical trial site operations, Deglise-Favre-Hawkinson (2015) studies capacity reservation and time windows for service, but focuses on the selection of which clinical trials to conduct subject to capacity. Turkcan et al. (2012) studies an optimization approach to planning as well as scheduling for chemotherapy infusion. They assume the desired series of care visits along a planned time profile is known, and their two-stage optimization model sequentially decomposes the planning and scheduling phases. Our paper differs in key ways, including the important feature of a network model (feature 1 above) and the stochastic itineraries of care (feature 3).

The work on integrated care systems is fairly limited. Hulshof et al. (2013) develops an intermediate horizon admission planning model that seeks "to achieve equitable access [delay] for patients, to meet production targets/to serve the strategically agreed number of patients, and to use resources efficiently." They optimize the system to meet throughput requirements, efficiency, and weights in the objective to prioritize the service of patients "at a particular stage in a particular care process." Their model allows a variety of resources and patient types, but a critical difference is that they assume deterministic arrivals and resource requirements, whereas our model allows for stochastic models for each of these.

Integrated outpatient care also has similarities to hospital inpatient scheduling problems. Elective patient admission scheduling research, including the studies of Adan et al. (2009), Chow et al. (2011), Bekker and Koeleman (2011), and Helm and Van Oyen (2014), has treated the optimization of elective admission schedules for stochastic flows through a network of inpatient hospital resources (e.g., wards). These studies, however, consider elective scheduling rather than capacity reservation, which is more appropriate for outpatient networks. The former sets the admission policies to achieve efficiency and low variability flow, and it does not focus on access delay under stochastic arrivals.

The concept of capacity reservation/allocation is present in some revenue management oriented research (e.g., Akkan 1997, Gupta and Liu 2008, Hsu and Wang 2001, Mula et al. 2006, Talluri and Van Ryzin 2006), but those models lack the features and complexity proposed in this work.

3 Model and System Dynamics

From here on, we will refer to our solution methodology and our research software instantiation as APT for "Access Planning Technology." APT's main output for managers of an outpatient care network is a template for planned capacity allocation. APT balances the tradeoffs between achieving (1) short access delays to a root appointment, (2) high utilization of clinicians' time, and (3) low probability of workload exceeding regular-time capacity. The main decision variables,

$$\Theta = \left(\Theta_t^{k,\tau}\right),$$

specificy the maximum number of patients of *class* $\tau \in \mathcal{C}(k)$ patients to be admitted on day $t \in \{1, \ldots, T\}$ to department/service $k \in \mathcal{K}$ for a root appointment of an itinerary. For modeling purposes, we will refer to the tuple (k, τ) as the patient *type*. In our examples, $\mathcal{K}$ is {GI, GIM, Neuro}. If $k =$ Neuro, then in our study $\mathcal{C}$(Neuro) is {Urgent, Non-Urgent}. Thus, for example, decision variable $\Theta_t^{\text{Neuro,Urgent}}$ is the maximum number of urgent patient root appointments allowed in Neurology on day t. Follow-up appointments and internal referrals are controlled indirectly, being scheduled into the remaining capacity left over after allocating Θ for all the root appointments.

There are three primary inputs to APT: (1) capacity of department k on day t measured in physician hours, denoted C_t^k, (2) exogenous demand, $X_t^{k,\tau}$, which is the random variable for the number of patients of class τ that request an appointment in department k on day t, and (3) downstream demand that is stochastically generate by each root appointment, which is described by a *stochastic location function*.

In the dynamics of our model, demand is either scheduled into the current day if capacity is available or carries over to the following day. To link this to an outpatient practice's actual process, a scheduler would receive a patient request to start a new itinerary. The scheduler would then sequentially check each future day of the scheduling template, Θ, for appointment slot availability for the appropriate patient type until a day with sufficient capacity is found and the patient is booked into that slot. The patient would then be informed, in real time, of the future appointment availability. Approximating current practice, we assume that patients are booked in a FCFS manner within patient type, which is the only possible mechanism given that there is no "queue" to choose from at the time a given patient calls. Our queueing model described below mimics the dynamics of this booking system.

On a given day t of our planning horizon, the class τ demand in service k can be split into: (i) the exogenous demand $X_t^{k,\tau}$ for a root appointment that is received on the current day t, and (ii) the "carryover demand" that represents all previously made requests that were not scheduled up to day t due to lack of template capacity. We refer to the combination of (i) and (ii) as the *Demand In Progress (DIP)* (similar to the concept Work In Progress (WIP) for queueing networks). The distribution of the DIP (Sec. 3.1) drives both the access delay (Sec. 3.2) and the total workload (Sec. 3.3).

The main modeling assumptions underlying the APT decision framework are as follows:

1. The exogenous demands $X_t^{k,\tau}$ are mutually independent and independent of all other inputs and decisions. In our case study, the arrivals form a cyclo-stationary process with a 5-day workweek as the system's period; e.g., successive Mondays are i.i.d., but have a different distribution than other days.

2. Within type (k, τ), patients are scheduled on a FCFS basis.

3. The total workload is assumed to be well approximated by a Normal distribution.

4. If the workload from downstream appointments exceeds capacity, the workload is served through overtime as opposed to being carried over into the future.

Collaborators at our partner institution indicated that the first two assumptions closely match what they observe in practice. Assumption 3 is empirically validated in our data in Sec. 3.3. Our partner organization also affirmed that, in their context, overtime is essentially unlimited. If there are patients scheduled for a particular day, the providers will stay until all have been served. It is extremely rare in their system that someone would have an appointment but not be seen. In the following sections, we characterize the DIP distribution and use it to analytically compute our key performance metrics:

1. **(M1)** Access delay (mean and service level) by patient type (k, τ).
2. **(M2-M3)** Mean and variance, respectively, of the resource utilization.
3. **(M4)** Expected amount (in hours) by which total workload exceeds capacity C_t^k.
4. **(M5)** Probability that the total workload exceeds capacity C_t^k.

3.1 Demand in Progress (DIP)

Let $D_t^{k,\tau}$ be the random variable that represents the amount of class τ DIP (number of patients) seeking an appointment in service k on a given day t of the planning horizon. The DIP accumulates to the next day, $t+1$, recursively:

$$D_{t+1}^{k,\tau} = X_{t+1}^{k,\tau} + \left[D_t^{k,\tau} - \Theta_t^{k,\tau}\right]^+,$$

where '+' denotes the positive part ($x^+ = \max(x, 0)$). For our application, we consider a cyclo-stationary system with a cycle of T business days (we consider T=5; Monday-Friday). Non-cyclostationary models are also possible within the framework. As a result, we can rewrite the above equation as follows:

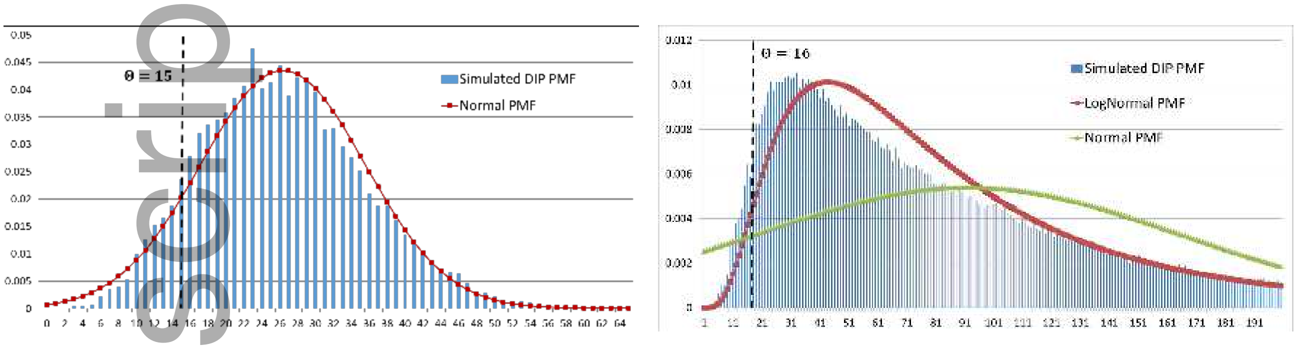
$$D_t^{k,\tau} = X_t^{k,\tau} + \beta_t^{k,\tau},\ t = 1, \ldots, T, \tag{1}$$

$$\beta_{t\oplus 1}^{k,\tau} = \left[D_t^{k,\tau} - \Theta_t^{k,\tau}\right]^+, \tag{2}$$

where $\beta_{t\oplus 1}^{k,\tau}$ is defined as the *carryover* demand from weekday t to the following weekday. The operator $\oplus$ is the modulo T operator: if $t = T$ then $t \oplus 1 = 1$. As we will see, $\beta_t^{k,\tau}$ drives the distribution of the access delay (Sec. 3.2). The workload distribution (Sec. 3.3) is driven by the the amount of DIP that is met on weekday t, denoted by $\alpha_t^{k,\tau}$:

$$\alpha_t^{k,\tau} = \min\{D_t^{k,\tau}, \Theta_t^{k,\tau}\} = D_t^{k,\tau} - \beta_{t\oplus 1}^{k,\tau},\ t = 1, \ldots, T. \tag{3}$$

Figure 3 displays some simulated DIP distributions (using the simulation described in Online Appendix C.6) for a single day of the week. In some cases, e.g., Fig. 3(a), DIP is well approximated by a Normal distribution. In other cases, when utilization is very high, e.g., Fig. 3(b), the DIP distribution has a heavier tail. A Normal distribution allows for an all-encompassing online optimization (Sec. 4), while for non-Normal DIP we use an iterative technique that combines optimization and simulation (Online Appendix D). For either case, we want to translate the set of stochastic, nonlinear equations (1)-(2) into a set of deterministic expressions that are linear in Θ (Sec. 4). In the next section, we describe how DIP can be used to calculate access delay.



(a) Lower Traffic (b) Heavier Traffic

Figure 3: Simulated DIP distributions (number of patients) for a single day, patient class, and department. (a) Lower traffic example: Monday DIP distribution with a mean of 26.36 and standard deviation of 9.16, which is well approximated by a Normal distribution (26.36, 9.16). (b) Heavier traffic example: Monday DIP distribution with a mean of 91.91 and standard deviation of 73.91, which is more closely approximated by a heavier tailed Log-Normal distribution (91.91, 73.91) than a Normal distribution (91.91, 73.91). In both examples, as is common, the DIP mean is greater than the template capacity Θ.

3.2 Access Delay: Metric M1

In this section, we develop analytical formulas for mean access delay and service level constraints on access delay. First, mean access delay for urgent patients can be formulated as:

$$\frac{\sum_{t=1}^{T} \mathbb{E}\left[\beta_t^{k,\text{Urgent}}\right]}{\sum_{t=1}^{T} \mathbb{E}\left[X_t^{k,\text{Urgent}}\right]}. \tag{4}$$

Here, $T^{-1}\sum_{t=1}^{T} \mathbb{E}\left[\beta_t^{k,\text{Urgent}}\right]$ represents the average number of urgent patients waiting per day (T is the length of the stationary cycle) (i.e., average queue length). $T^{-1}\sum_{t=1}^{T} \mathbb{E}\left[X_t^{k,\text{Urgent}}\right]$ represents the average access demand for appointments (i.e., average arrival rate). Using Little's Law, we get the average delay to obtain an appointment (i.e., average waiting time) by dividing the two quantities: long-run average time in queue (access delay) equals the long-run average number of patients in the queue (overflow demand) divided by the long-run average arrival rate.

Next we formulate service level constraints on access delay, which we define as a limit on the fraction of patients whose delay to obtain a root appointment exceeds a specified number of days. We let the fraction and number of days be patient type-specific and selected a priori by the user (manager). To do so, we define each service level constraint as a tuple, $(p_n^{k,\tau}, TFAV_n^{k,\tau})$, which indicates that $p_n^{k,\tau}$ is the upper bound on the percentage of class τ patients in service k that will exceed a *time to first available* *visit* (i.e., access delay) of $TFAV_n^{k,\tau}$ days. The subscript n allows us to set multiple bounds for each patient type. For example, we may want the first service level constraint ($n = 1$) to be $(0.2, 4)$, which means that 20% ($p_1^{k,\tau} = 0.2$) of type (k, τ) patients get an appointment within 4 days ($TFAV_1^{k,\tau} = 4$). We might also want to include a second service level constraint ($n = 2$) for type (k, τ) patients as $(0.5, 7)$, which means that 50% ($p_2^{k,\tau} = 0.5$) of type (k, τ) patients will get an appointment within 7 days ($TFAV_2^{k,\tau} = 7$). This approach actually allows us to have control over the distribution of access delay, as demonstrated in Sec. 5.3.2.

To capture these service level metrics, we begin by defining $\delta_{t,n}^{k,\tau}$ as the total number of open slots left in our template from day t up to day $t + TFAV_n^{k,\tau}$ after all demand prior to day t has been scheduled:

$$\delta_{t,n}^{k,\tau} = \left[\left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t \oplus l}^{k,\tau} \right) - \beta_t^{k,\tau} \right]^+ . \tag{5}$$

This is the positive difference of (i) the total number of type (k, τ) slots from day t to day $t + TFAV_n^{k,\tau}$ (i.e., the access delay limit for patients requesting an appointment on or before day t) minus (ii) the number of type (k, τ) carryovers to day t (i.e., the number of patients that requested an appointment on or before day t and have yet to be assigned an appointment). $\delta_{t,n}^{k,\tau}$ therefore represents the number of slots remaining before the TFAV deadline that can be used to satisfy the day t demand.

The expected fraction of class τ patients requesting an appointment in service k on day t (maintaining FCFS) that exceed $TFAV_n^{k,\tau}$ days of delay to obtain a root appointment is denoted by $G_{t,n}^{k,\tau}$:

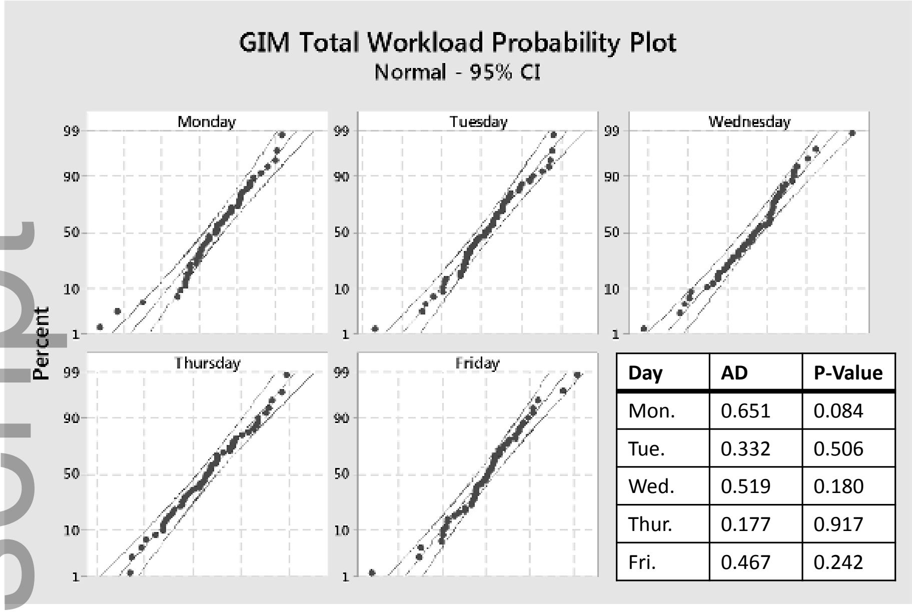
$$G_{t,n}^{k,\tau} = \mathbb{E}\left[\frac{\left(X_t^{k,\tau} - \delta_{t,n}^{k,\tau} \right)^+}{X_t^{k,\tau}} \right] . \tag{6}$$

Note that $X_t^{k,\tau}$ is independent of $\delta_{t,n}^{k,\tau}$. The service level constraints are of the form $G_{t,n}^{k,\tau} \le p_n^{k,\tau}$.

3.3 Linearity of the Clinic Workload Process: Metrics M2 and M3

We model the workload using an offered load approach, leveraging a *stochastic location function* to capture the downstream resource requirements generated by each root appointment (e.g., Leung et al. 1994). We then approximate the resulting Poisson-distributed offered load by a Normal distribution. This approximation is validated in our data of total workload in physician hours. For example, Figure 4 displays Normal probability plots, 95% confidence bands, and Anderson-Darling test statistics and p-values (the higher, the better) for the historical GIM total workload by day. Note that, as desired under a Normal approximation, the data points form nearly straight lines. Other departments are similar.

The Normal distribution is fully specified through its mean and variance. In this section, we show (Theorem 3.1) that the workload mean can be expressed linearly in the mean amount of DIP met, $\mathbb{E}[\alpha]$,



Day	AD	P-Value
Mon.	0.651	0.084
Tue.	0.332	0.506
Wed.	0.519	0.180
Thur.	0.177	0.917
Fri.	0.467	0.242

Figure 4: Normal probability plots of GIM's historical total workload (in physician hours), by day of week. The Anderson-Darling (AD) test statistics are presented along with the associated p-values, where the null hypothesis is that the data follow a Normal distribution. At a 5% level of significance (95% confidence), there is insufficient evidence to reject the null hypothesis since the p-values are above .05. This Minitab® 17 output also displays the 95% confidence interval (CI) bands.

which is the only quantity that depends on Θ. Likewise, the workload variance only depends on Θ through α; and, we show (Theorem 3.2) that the workload variance is linear in $\mathbb{E}[\alpha]$ and $\mathrm{Var}[\alpha]$.

First, we mathematically define the patient's path through the network of specialist services. Since a patient could have downstream appointments at multiple medical specialties on a single day, we need to consider a vector state space for the stochastic location process. Let this vector state space be $\mathcal{S}^0 = \{[a_1, a_2, \ldots, a_{|\mathcal{K}|}] : a_k \in \mathbb{Z}^+, \forall k \in \mathcal{K}\}$, where a_k is the number of time slots the patient requires in service k. We let the full state space be $\mathcal{S} = \mathcal{S}^0 \cup \{\Delta\}$, where Δ represents that the patient has no appointments (e.g., has returned home, has not yet become a downstream patient, or has no visits on a given day within his/her itinerary). The $\mathcal{S}$-valued *stochastic location function* denoted by $L_{t_1}^{k_1,\tau}(t)$ represents the number of appointment slots needed at time t during a care episode for a patient of class τ that started her itinerary with a root appointment in service k_1 at time t_1. We define the *resource probabilities*, r, as follows:

$$r_{t_1}^{k_1,\tau,k}(m, t - t_1) = \mathbb{P}(L_{t_1}^{k_1,\tau}(t) \cdot \mathbf{e}_k = m),$$

where $\mathbf{e}_k$ is a column vector with all 0's and a 1 in the k^{th} row. Then, $r_{t_1}^{k_1,\tau,k}(m, t)$ is the probability that a class τ root appointment in department k_1 on day t_1 will result in m downstream appointment slots t days later in department k. These resource probabilities are calibrated from historical data; for an example see Table 2 in Sec. 5.1. While we may jointly optimize the templates of all departments in

$\mathcal{K}$, there may be other outside departments with static, uncontrolled templates that still refer patients to the departments in $\mathcal{K}$. That is, $k_1 \in \mathcal{K}'$ where, for example, $\mathcal{K}' \equiv \mathcal{K} \cup \{\text{Other}\}$.

Define W_t^k as the total workload (root appointments and downstream visits) in service k on day t, measured in terms of physician hours. The next two theorems show that the first two moments of W_t^k (i.e., $\mathbb{E}[W_t^k]$ and $\text{Var}[W_t^k]$) can be expressed linearly in the mean and variance of the scheduled demand for day t in service k. Let M_k be the maximum number of appointment time slots a patient can require within a day in specialty k, and let s_k be the time (in physician hours) per slot in specialty k. We assume that s_k is a deterministic input.

Theorem 3.1. *The steady state mean offered workload in service k on day t (under the capacity reservation plan Θ) can be computed as:*

$$\mathbb{E}[W_t^k] = \sum_{k_1 \in \mathcal{K}'} \sum_{\tau \in \mathcal{C}(k_1)} \sum_{t_1=1}^{T} \mathbb{E}[\alpha_{t_1}^{k_1,\tau}] \cdot \sum_{j=0}^{\infty} \sum_{m=1}^{M_k} m \cdot r_{t_1}^{k_1,\tau,k}(m, t-t_1+jT) \cdot s_k. \tag{7}$$

Theorem 3.2. *The steady state variance of the offered workload in specialty k on day t of our steady state planning horizon is given by*

$$\begin{aligned}\text{Var}[W_t^k] = \sum_{k_1 \in \mathcal{K}'} \sum_{\tau \in \mathcal{C}(k_1)} \sum_{t_1=1}^{T} \sum_{j=0}^{\infty} \Bigg[& \text{Var}\left[\alpha_{t_1}^{k_1,\tau}\right] \left(\sum_{m=0}^{M_k} m \cdot r_{t_1}^{k_1,\tau,k}(m, t-t_1+jT) \cdot s_k \right)^2 \\ & + \mathbb{E}\left[\alpha_{t_1}^{k_1,\tau}\right] \cdot \sum_{m=0}^{M_k} \Bigg(m^2 \cdot s_k^2 \cdot r_{t_1}^{k_1,\tau,k}(m, t-t_1+jT)\left(1 - r_{t_1}^{k_1,\tau,k}(m, t-t_1+jT)\right) \\ & - \sum_{m<q\le M_k} 2mq \cdot s_k^2 \cdot r_{t_1}^{k_1,\tau,k}(m, t-t_1+jT) \cdot r_{t_1}^{k_1,\tau,k}(q, t-t_1+jT) \Bigg) \Bigg].\end{aligned} \tag{8}$$

Proofs for all theorems and propositions are presented in Online Appendix A. According to Theorem 3.1, the workload mean is linear in $\mathbb{E}[\alpha_{t_1}^{k_1,\tau}]$, which are the only quantities that depend on the control Θ. According to Theorem 3.2, the workload variance depends linearly on $\mathbb{E}[\alpha_{t_1}^{k_1,\tau}]$ and $\text{Var}[\alpha_{t_1}^{k_1,\tau}]$, which are the only quantities that depend on Θ. A challenge then is to express the mean and variance of $\alpha_{t_1}^{k_1,\tau}$ linearly in Θ. Once successful (Sec. 4.3), since utilization is the ratio of total workload to a deterministic capacity, the fact that metrics M2-M3 can be expressed linearly in Θ easily follows.

4 Mixed Integer Program

The equations that specify the system dynamics (Eqs. (1)–(3)) and performance metrics (Eqs. (6), (7), and (8)) are both stochastic and nonlinear in the decision variables Θ. In this section, we transform these equations into a set of expressions that are both deterministic and linear in Θ. This allows us to formulate an MIP that can be solved by commercial solvers like IBM CPLEX.

4.1 Linear Formulation of the System Dynamics

The first step in formulating the deterministic, linear MIP is to discretize the DIP distribution. As discussed above, the DIP distribution may be well approximated as Normal (Fig. 3(a)). However, when utilization is very high, DIP tends to have a heavier tail (Fig. 3(b)). In this section, we restrict attention to the Normal case. For non-Normal DIP, we use the technique presented in Online Appendix D to adjust the approach presented here. In either case, simulation may be used to verify that the metrics calculated within the optimization are close to the true, simulated metrics for the output template.

We first discretize the DIP distribution through an approximation based on Riemann integration. Let $\mathcal{I} \equiv \{1, 2, \ldots, I\}$ be an index that creates a discrete grid with $I+1$ sections. First, grid point i is located at $m(i)$ standard deviations above the DIP mean. Define $\Psi(i)$ to be the probability mass contained within the interval between the $(i-1)^{st}$ and i^{th} grid points: $(\mu + m(i-1)\sigma, \mu + m(i)\sigma]$ for mean μ and standard deviation σ. For an example, see the grid in Table 1. We then interpret $\Psi(i)$ as the probability that the realized value of $D_t^{k,\tau}$ lies within $(\mu + m(i-1)\sigma, \mu + m(i)\sigma]$. As in Riemann integration, the grid does not need to be linear. It is important to note that, due to the fact that any Normal distribution can be standardized, the probability masses, Ψ, can be calculated off line using a Standard Normal distribution and enter the optimization as inputs.

$i \in \mathcal{I}$	1	2	3	4	5	6	7	8	9	10	11	12	13
$m(i)$	-3.1	-1.8	-1.2	-0.6	-0.2	-0.1	0.0	0.1	0.2	0.6	1.2	1.8	3.1
$\Psi(i)$	.036	.079	.159	.146	.039	.040	.040	.039	.146	.159	.079	.035	.001

Table 1: Sample grid mapping along with p.m.f. Ψ.

We now define the variable $D_t^{k,\tau}(i)$ as the realization of the DIP at $m(i)$ standard deviations above the mean, where the DIP mean and standard deviation are calculated within the optimization:

$$D_t^{k,\tau}(i) = \mathbb{E}[D_t^{k,\tau}] + m(i)\sqrt{\mathrm{Var}[D_t^{k,\tau}]}. \tag{9}$$

To apply MIP optimization techniques, we need to express $D_t^{k,\tau}(i)$ linearly in Θ. To do so, we linearize $\mathbb{E}[D_t^{k,\tau}]$ and $\sqrt{\mathrm{Var}[D_t^{k,\tau}]}$ below. To linearize this standard deviation, we will first show that $\mathrm{Var}[D_t^{k,\tau}]$ can be expressed linearly in Θ. Then, we approximate the standard deviation by applying one step of Newton's method:

$$\sqrt{\mathrm{Var}[D_t^{k,\tau}]} \approx \frac{1}{2}(\mathrm{Var}[D_t^{k,\tau}]/\hat{D}_t^{k,\tau} + \hat{D}_t^{k,\tau}), \tag{10}$$

where $\hat{D}_t^{\tau}$ denotes a "guess" for the true value of $\sqrt{\mathrm{Var}[D_t^{k,\tau}]}$. The $\hat{D}_t^{\tau}$ are inputs to the MIP. In practice, we may calibrate $\hat{D}_t^{\tau}$ using historical data. Alternatively, we may fix a (non-optimized) template and simulate the DIP standard deviation to obtain $\hat{D}_t^{\tau}$. In previous work, Helm and Van Oyen (2016) showed that a similar Newton's method approximation is highly effective for modeling offered workloads, and that the approximation is robust to deviations of the fixed/historical estimate (in our case, $\hat{D}^{\tau}_t$) from the true standard deviation. If we can show that both $\mathbb{E}[D^{k,\tau}_t]$ and $\mathrm{Var}[D^{k,\tau}_t]$ are linear in Θ, then, since $\hat{D}^{\tau}_t$ is constant, the approximation of $\sqrt{\mathrm{Var}[D^{k,\tau}_t]}$ will also be linear in Θ. Next, we show (Theorem 4.1) that $D^{k,\tau}_t(i)$ can be expressed linearly in Θ. Recall from Eqs. (1)–(2) the following DIP recursive relationship: $D^{k,\tau}_{t\oplus1} = X^{k,\tau}_{t\oplus1} + \beta^{k,\tau}_{t\oplus1}$, where $\beta^{k,\tau}_{t\oplus1} = [D^{k,\tau}_t - \Theta^{k,\tau}_t]^+$. In order to linearize $D^{k,\tau}_{t\oplus1}(i)$, we introduce additional "helper" decision variables $\beta^{k,\tau}_{t\oplus1}(i)$ defined as follows:

$$\beta^{k,\tau}_{t\oplus1}(i) = [D^{k,\tau}_t(i) - \Theta^{k,\tau}_t]^+.$$

We interpret $\beta^{k,\tau}_{t\oplus1}(i)$ as the realized value of $\beta^{k,\tau}_{t\oplus1}$ conditioned on DIP equaling $D^{k,\tau}_t(i)$. In order to linearize $\beta^{k,\tau}_{t\oplus1}(i)$, we define $y^{k,\tau}_t(i,l)$ as a binary variable that equals 1 when $D^{k,\tau}_t(i) - \Theta^{k,\tau}_t > l$, and 0 otherwise; $l = 0, 1, \ldots$. (Technically, for Normal DIP, $l \in [0,\infty)$, but we discretize l since DIP is discrete in practice.) The following constraints will guarantee that this definition is satisfied:

$$-M \cdot \left(1 - y^{k,\tau}_t(i,l)\right) \le D^{k,\tau}_t(i) - \Theta^{k,\tau}_t - l, \tag{11}$$

$$D^{k,\tau}_t(i) - \Theta^{k,\tau}_t - l \le M \cdot y^{k,\tau}_t(i,l), \tag{12}$$

where M is a sufficiently large number. The following constraints linearize $\beta^{k,\tau}_{t\oplus1}(i)$:

$$\beta^{k,\tau}_{t\oplus1}(i) \ge D^{k,\tau}_t(i) - \Theta^{k,\tau}_t, \tag{13}$$

$$\beta^{k,\tau}_{t\oplus1}(i) \le D^{k,\tau}_t(i) - \Theta^{k,\tau}_t + M \cdot \left(1 - y^{k,\tau}_t(i,0)\right). \tag{14}$$

We can now approximate the mean of $\beta^{k,\tau}_{t\oplus1}$ by $\overline{\beta}^{k,\tau}_{t\oplus1}$ (using the bar to denote expectation) as follows:

$$\overline{\beta}^{k,\tau}_{t\oplus1} = \sum_{i\in\mathcal{I}} \beta^{k,\tau}_{t\oplus1}(i)\Psi(i).$$

Since $\Psi(i)$ is the probability mass for the conditional value $\beta^{k,\tau}_{t\oplus1}(i)$, it follows from the Law of Total Expectation that $\overline{\beta}^{k,\tau}_{t\oplus1}$ is an expected value. In the Riemann limit, as $I \to \infty$ and $\Psi(i) \to 0$, $\overline{\beta}^{k,\tau}_{t\oplus1}$ converges to $\mathbb{E}[\beta^{k,\tau}_{t\oplus1}]$. Note that since expectation is a linear operator, if the $\beta^{k,\tau}_{t\oplus1}(i)$ can be expressed linearly in Θ, then $\overline{\beta}^{k,\tau}_{t\oplus1}$ can also be expressed linearly in Θ. Similarly, define $\tilde{\beta}^{k,\tau}_{t\oplus1}$ as a value that converges, in the Riemann limit, to $\mathrm{Var}[\beta^{k,\tau}_{t\oplus1}]$. (The '$\sim$' denotes variance throughout.) The following proposition demonstrates that $\tilde{\beta}^{k,\tau}_{t\oplus1}$ can be linearized.

Proposition 4.1. *Using the previously defined binary variables $y^{k,\tau}_t(i,l)$, the overflow demand variance can be expressed linearly as follows:*

$$\tilde{\beta}^{k,\tau}_{t\oplus1} = \sum_{l=0}^{\infty}\left[(2l+1)\cdot\sum_{i\in\mathcal{I}} y^{k,\tau}_t(i,l)\cdot\Psi(i)\right] - \sum_{(l_1,l_2)\in(\mathbb{Z}^+)^2}\ \sum_{(i_1,i_2)\in\mathcal{I}^2} z^{k,\tau}_t(i_1,i_2,l_1,l_2)\cdot\Psi(i_1)\cdot\Psi(i_2),$$

where the $z_t^{k,\tau}(i_1,i_2,l_1,l_2)$ are binary variables that satisfy the following constraints:

$$z_t^{k,\tau}(i_1,i_2,l_1,l_2) \le y_t^{k,\tau}(i_1,l_1), \tag{15}$$

$$z_t^{k,\tau}(i_1,i_2,l_1,l_2) \le y_t^{k,\tau}(i_2,l_2), \tag{16}$$

$$z_t^{k,\tau}(i_1,i_2,l_1,l_2) \ge y_t^{k,\tau}(i_1,l_1) + y_t^{k,\tau}(i_2,l_2) - 1. \tag{17}$$

Theorem 4.1. *Within an MIP, variables $D_t^{k,\tau}(i)$ and $\beta_t^{k,\tau}(i)$, $i \in \mathcal{I}$, and $\overline{\beta}_t^{k,\tau}$ and $\tilde{\beta}_t^{k,\tau}$, can be expressed linearly in Θ using additional binary variables $y_t^{k,\tau}(i,l)$, $i \in \mathcal{I}$, $l \in \mathbb{Z}^+$, and $z_t^{k,\tau}(i_1,i_2,l_1,l_2)$, $i_1 \in \mathcal{I}$, $i_2 \in \mathcal{I}$, $l_1 \in \mathbb{Z}^+$, $l_2 \in \mathbb{Z}^+$; $t = 1,2,\ldots,T$.*

Even with linearization, tractability can be an issue, as the number of binary variables $z_t^{k,\tau}(i_1,i_2,$ $l_1,l_2)$ needed to linearize the carryover variance (Proposition 4.1) can become very large. To overcome this challenge, in Online Appendix B we propose a linear approximation of the carryover variance that works very well in practice, and is validated in Sec. 5.2.

Next, we use these system dynamics to formulate the access delay and workload/overtime metrics linearly in Θ. A full MIP formulation is presented in Online Appendix C. There are additional constraints that appear in the formulation for practical reasons: (1) We add variables to ensure that the DIP distribution is nonnegative. (2) We add cuts (orderings) for the binary variables that tend to speed up computation. Finally, we note that we truncate infinite sums, as detailed in Online Appendix C.

4.2 Linear Transformation of the Access Delay: Metric M1

In this section, we formulate our service level constraints on access delay. Specifically, we linearly approximate with respect to our decision variables, Θ, the expected fraction, $G_{t,n}^{k,\tau}$, of class τ patients requesting an appointment in service k on day t that exceed $TFAV_n^{k,\tau}$ days of delay to obtain a root appointment. Recall (Eq. (6)) that $G_{t,n}^{k,\tau}$ is a function of $\left(X_t^{k,\tau} - \delta_{t,n}^{k,\tau}\right)^+$. Because the exogenous demand at t is independent of past demand/decisions, $X_t^{k,\tau}$ and $\delta_{t,n}^{k,\tau}$ are independent. We condition $\delta_{t,n}^{k,\tau}$ on the event that the total demand is $D_{t\ominus 1}^{k,\tau}(i)$ on day $t-1$ (see Eq. (11)), and obtain:

$$\delta_{t,n}^{k,\tau}(i) = \left(\left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t\oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i)\right)^+, \forall i \in \mathcal{I}, \forall n \in \mathcal{N}, \tag{18}$$

where $\mathcal{N} \equiv \{1,2,\ldots,N\}$ and N is the total number of $TFAV_n^{k,\tau}$ targets.

To capture the joint distribution between $X_t^{k,\tau}$ and $\delta_{t,n}^{k,\tau}$ (which is simplified by their independence), we define $\gamma_{t,n}^{k,\tau}(i,j)$ as the percentage of class τ patients requesting an appointment in service k on day t that exceed $TFAV_n^{k,\tau}$ days of waiting for their appointment given that (i) there are j class τ requests in service k on day t, and (ii) there are $\delta_t^{k,\tau}(i)$ class τ slots in the template for service k prior to the

TFAV deadline after all demand prior to day t has been scheduled. Then, $\gamma_{t,n}^{k,\tau}(i,j)$ can be expressed as follows:

$$\gamma_{t,n}^{k,\tau}(i,j) = \frac{\left(j - \delta_{t,n}^{k,\tau}(i)\right)^{+}}{j} = \left(1 - \frac{\delta_{t,n}^{k,\tau}(i)}{j}\right)^{+}, \forall i \in \mathcal{I}, \forall j \in \mathcal{J}, \forall n \in \mathcal{N}, \tag{19}$$

where the set $\mathcal{J} \subseteq \mathbb{Z}^{+}$ represents the sample space of all the $X_t^{k,\tau}$ random variables (excluding the outcome equal to 0). Thus, the access delay metric is given by:

$$G_{t,n}^{k,\tau} = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \gamma_{t,n}^{k,\tau}(i,j) \cdot \Psi(i) \cdot \mathbb{P}(X_t^{k,\tau} = j), \forall n \in \mathcal{N}, \tag{20}$$

where $\mathbb{P}(X_t^{k,\tau} = j)$ is an input calibrated using historical data.

In order to have linear constraints, we need to alter Eqs. (18) and (19). In the optimization, there is an incentive to keep $\gamma_{t,n}^{k,\tau}(i,j)$ small to meet the access delay constraints (or because we are minimizing it in our objective). This allows us to replace Eq. (19) with the following:

$$\gamma_{t,n}^{k,\tau}(i,j) \geq \left(1 - \frac{\delta_{t,n}^{k,\tau}(i)}{j}\right), \ \gamma_{t,n}^{k,\tau}(i,j) \geq 0, \forall i \in \mathcal{I}, \forall j \in \mathcal{I}, \forall n \in \mathcal{N}. \tag{21}$$

The same cannot be said for Eq. (18), since the optimization has the incentive to increase $\delta_{t,n}^{k,\tau}(i)$ in order to get a smaller $\gamma_{t,n}^{k,\tau}(i,j)$ that will meet the access delay constraints. Therefore, we introduce a binary variable $x_{t,n}^{k,\tau}(i)$ that equals 1 when $\left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t \oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i) \geq 0$ and equals 0 otherwise. The following constraints will ensure that the $x_{t,n}^{k,\tau}(i)$ take on the correct values:

$$-M \cdot (1 - x_{t,n}^{k,\tau}(i)) \leq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t \oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i), \forall i \in \mathcal{I}, \forall n \in \mathcal{N}, \tag{22}$$

$$M \cdot x_{t,n}^{k,\tau}(i) \geq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t \oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i), \forall i \in \mathcal{I}, \forall n \in \mathcal{N}. \tag{23}$$

We now have a set of linear inequalities equivalent to Eq. (18):

$$\delta_{t,n}^{k,\tau}(i) \geq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t \oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i), \forall i \in \mathcal{I}, \forall n \in \mathcal{N}, \tag{24}$$

$$\delta_{t,n}^{k,\tau}(i) \leq \left(\sum_{l=0}^{TFAV_n^{k,\tau}} \Theta_{t \oplus l}^{k,\tau}\right) - \beta_t^{k,\tau}(i) + M \cdot (1 - x_{t,n}^{k,\tau}(i)), \forall i \in \mathcal{I}, \forall n \in \mathcal{N}, \tag{25}$$

$$\delta_{t,n}^{k,\tau}(i) \leq M \cdot x_{t,n}^{k,\tau}(i), \forall i \in \mathcal{I}, \forall n \in \mathcal{N}. \tag{26}$$

4.3 Linear Transformation of Excess Workload and Overtime: Metrics M4 and M5

Metric M4: Expected Overtime. In this section, we use the total workload, W_t^k, to calculate the amount of overtime due to insufficient capacity in specialty k on day t. Recall that W_t^k is Normally distributed. Under this assumption, the distribution of W_t^k is specified by its mean and variance. In Theorem 3.1, note that the workload mean depends on the decision variables, Θ, only through $\mathbb{E}[\alpha_t^{k,\tau}]$; and, in Theorem 3.2, the variance depends on Θ through $\mathbb{E}[\alpha_t^{k,\tau}]$ and $\mathrm{Var}[\alpha_t^{k,\tau}]$. Moreover, the workload mean is linear in $\mathbb{E}[\alpha_t^{k,\tau}]$, and the workload variance is linear in $\mathbb{E}[\alpha_t^{k,\tau}]$ and $\mathrm{Var}[\alpha_t^{k,\tau}]$. To show that these workload moments can be expressed linearly in Θ, we then only need to express $\mathbb{E}[\alpha_t^{k,\tau}]$ and $\mathrm{Var}[\alpha_t^{k,\tau}]$ linearly in Θ. For $\mathrm{Var}[\alpha_t^{k,\tau}]$, we use the linear expression for $\tilde{\alpha}_t^{k,\tau}$ in Eq. (36) of Online Appendix B. For the mean, denoted by $\overline{\alpha}_t^{k,\tau}$, the following linear expression follows from Eqs. (1) and (3) and the fact that expectation is a linear operator:

$$\overline{\alpha}_t^{k,\tau} = \mathbb{E}[D_t^{k,\tau}] - \overline{\beta}_{t\oplus 1}^{k,\tau} = \mathbb{E}[X_t^{k,\tau}] + \overline{\beta}_t^{k,\tau} - \overline{\beta}_{t\oplus 1}^{k,\tau}. \tag{27}$$

Similar to the discretization of the DIP distribution in Sec. 4.1, we discretize the workload distribution using a grid approximation. For grid point i, the possible workload realization is $\overline{W}_t^k + m(i)\sqrt{\tilde{W}_t^k}$, and the associated probability mass is $\Psi(i)$. The mean, $\overline{W}_t^k$, and variance, $\tilde{W}_t^k$, are calculated using Theorems 3.1 and 3.2, respectively, with the linear expressions $\overline{\alpha}_t^{k,\tau}$ and $\tilde{\alpha}_t^{k,\tau}$. We capture the realization of overtime hours at grid point i using a variable $O_t^k(i)$ as follows:

$$O_t^k(i) = \left(\overline{W}_t^k + m(i)\cdot\sqrt{\tilde{W}_t^k} - C_t^k\right)^+ \approx \left(\overline{W}_t^k + \frac{1}{2}m(i)\cdot\left(\frac{\tilde{W}_t^k}{\hat{W}_t^k} + \hat{W}_t^k\right) - C_t^k\right)^+, \tag{28}$$

where C_t^k is the total capacity of specialty k on day t, and $\hat{W}_t^k$ is the initial guess for the standard deviation of the total workload on day t in specialty k. The approximation follows from the one-step Newton's method approximation detailed in Sec. 4.1. Then, the expected overtime hours, denoted $\overline{O}_t^k$, are calculated as follows:

$$\overline{O}_t^k = \sum_{i\in\mathcal{I}} O_t^k(i)\Psi(i). \tag{29}$$

Metric M5: Probability of Overtime. In addition to constraining the expected overtime amount (metric M4), we can also limit the probability of exceeding service k's capacity on a given day by some amount q_t^k (metric M5). First, we select the smallest $i^* \in \mathcal{I}$ such that $1 - \Phi(m(i^*+1)) \leq q_t^k$, where $\Phi(\cdot)$ is the Standard Normal c.d.f. Then, we can constrain the workload level at this grid point i^* to be less than or equal to capacity:

$$\overline{W}_t^k + \frac{1}{2}m(i^*)\cdot\left(\frac{\tilde{W}_t^k}{\hat{W}_t^k} + \hat{W}_t^k\right) \leq C_t^k. \tag{30}$$

5 Numerical Case Study

In this section, we present a case study based on a collaboration with several services at the same healthcare organization. This case study is comprised of several sub-studies, each designed to solve different managerial challenges faced by our industry partner. These studies were carried out over the course of several years through numerous interactions and iterations, and months of on-site work with our industry partner. We begin by describing the data that was available to us and how we used this data to validate our analytical approximations of our key metrics. Our first study analyzes the impact of template design and patient mix on access delay for both urgent and non-urgent patients. Our second study focuses on designing templates to accommodate the business need of increasing the volume of one particular patient type while still maintaining a high level of patient access.

5.1 Data

We obtained one calendar year of data containing the following data items for three services (GI, GIM, and Neuro) by patient class and by day: (1) capacity data, (2) histories of downstream appointments generated from a root appointment, (3) demands for new root appointments, and (4) internal referral workload to each service.

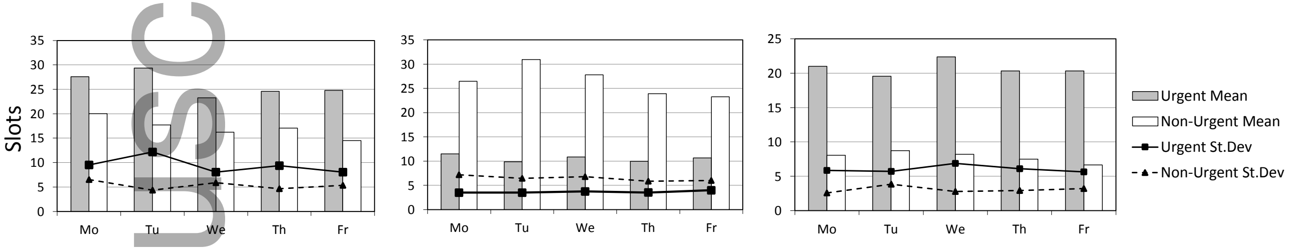
The capacity data indicate how many slots are available in each service by appointment type and by day, from which we can calculate the total physician hours available by day. These data are summarized for the three services in Figure 2 of Sec. 2. The data regarding type, timing, quantity, and length (how many time slots and how many minutes they take) of downstream appointments was used to estimate the p.m.f.'s of the stochastic location functions defined in Sec. 3.3. An example of three location probability matrices for urgent GI patients either returning to GI for follow-up appointments or being *internally referred* to Neurology or GIM is shown in Table 2. The rows indicate days after the root appointment, and the columns indicate the number of appointment slots required on that day in GI (left table), Neurology (middle table), and GIM (right table). For example, row 3 of the GI to GI matrix indicates that two days after a root appointment in GI, 2 appointment slots are scheduled for a follow-up in GI with probability 0.11 and zero slots with probability 0.89. Also two days after a root appointment, Neuro requires 4 slots with a probability of 0.02 and zero otherwise and GIM requires 2 slots with probability of 0.01 and zero otherwise.

The empirical distribution of new root appointment requests was generated from historical data on new requests for each service. The mean and standard deviation of the number of slots requested by each patient class (urgent vs. non-urgent) and day of week in each service are given in Fig. 5.

Finally, we used data on historical internal referral workloads from all other services (not just GI, GIM, and Neurology). The amount of internal referrals typically increases throughout the week, with Monday having the lowest level and Thursday and Friday having the highest amount of internal referrals. Table 3 shows the percent of total demand at each service (on average) that was made up of internal referrals. As can be seen from the percentages, planning for the impact of internal referrals from the

Table 2: Example of the location function for urgent GI patients. Rows are days after the root appointment, and columns are the number of appointment slots required on that day. GI to GI follow-up appointments are to the left, GI to Neuro internal referral is in the middle, and GI to GIM internal referral is on the right.

	GI to GI (Follow Up)							GI to Neuro					GI to GIM		
Days After	Number of Appt Slots							# Appt Slots					# Appt Slots		
Root Appt	0	1	2	3	4	5	6	0	1	2	3	4	0	1	2
0	0%	0%	0%	0%	55%	3%	42%	99%	0%	0%	0%	1%	100%	0%	0%
1	85%	0%	15%	0%	0%	0%	0%	97%	0%	0%	0%	3%	98%	0%	2%
2	89%	0%	11%	0%	0%	0%	0%	98%	0%	0%	0%	2%	99%	0%	1%
3	89%	0%	11%	0%	0%	0%	0%	99%	0%	0%	0%	1%	99%	0%	1%
4	91%	0%	9%	0%	0%	0%	0%	100%	0%	0%	0%	0%	99%	1%	0%
5	98%	0%	2%	0%	0%	0%	0%	100%	0%	0%	0%	0%	100%	0%	0%



(a) GIM (b) GI (c) Neurology

Figure 5: Mean and standard deviation of the empirical demand (number of appointment slots) for new root appointment requests for GIM, GI, and Neurology services.

aggregated departments of the network can be very important.

Table 3: Percent of total demand (on average) that is comprised of internal referrals for GI, GIM, and Neurology.

Svc \ Day	Mo	Tu	We	Th	Fr
GI	23%	28%	34%	41%	49%
GIM	38%	42%	45%	46%	47%
Neuro	13%	21%	22%	22%	21%

5.2 Simulation Validation of Analytical Approximations and Performance Metrics

In this section, we employ the discrete event simulation described in Online Appendix C.6 to validate the novel analytical approximations that make the optimization tractable. The simulation provides an accurate testbed to perform this validation, since it makes no approximations and instead directly models the dynamics of the system from the available data. Using the optimized template from Sec. 5.3, the simulation computes the performance of the metrics that are critical to the institution, which we compare with those from the analytical approximation.

Fig. 6 shows that the complementary c.d.f.'s of access delay computed via the analytical approximations are indeed very close to the results from the simulation model. Observe that the analytical approximations are shown to perform very well in predicting the c.d.f. of the access delay, not just the mean or variance. The absolute percentage errors (APEs) of our estimations are summarized in Table

4. Observe that, even at the distribution level, these errors are quite small even in the worst cases.

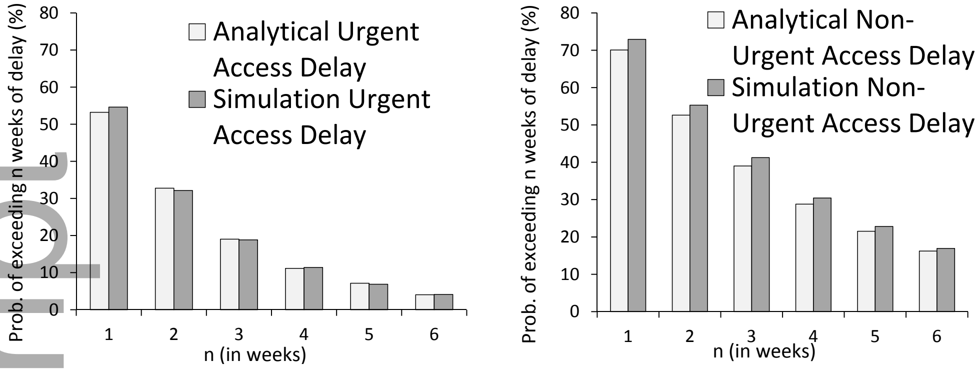


Figure 6: Comparison of the complementary c.d.f.'s for access delay (weeks) of the analytical model versus the simulation for (Left) urgent cases and (Right) non-urgent cases.

n (in weeks)	1	2	3	4	5	6
Urgent Access Delay APE (%)	2.58	1.91	1.01	2.18	3.98	1.95
Non-Urgent Access Delay APE (%)	3.90	4.84	5.47	5.39	5.67	4.16

Table 4: Absolute percentage error of the expected percentage of urgent and non-urgent patients exceeding n weeks of access delay.

Fig. 7 and Table 5 show the validation results for the mean and standard deviation approximations for clinic workloads. It is clear that the error in the analytical approximation of workload is likewise very small. We believe one reason this approximation is so accurate lies in the fact that the system we study from our partner organization tends to run under a heavy load, which makes daily workloads behave more predictably.

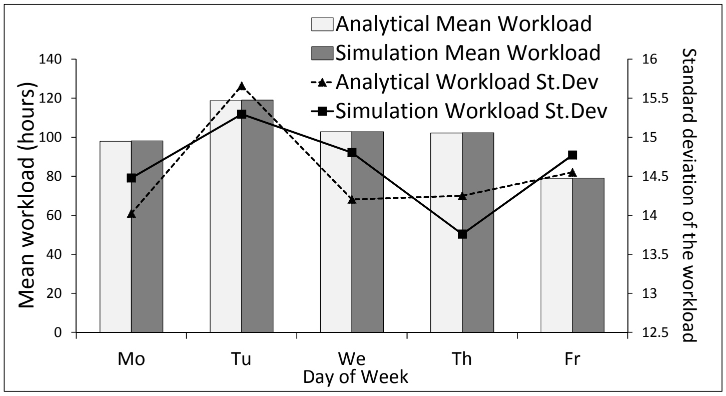


Figure 7: Comparison of the means and standard deviations of workload by day of week for the analytical model versus the simulation.

5.3 Improving Access

First, we consider the case where management has target performance levels and needs a template to achieve those targets. As a proof of concept, we present one particular scenario that was of strong

Day of Week	Mo	Tu	We	Th	Fr
Mean Workload APE (%)	0.29	0.28	0.04	0.15	0.36
Workload St.Dev. APE (%)	3.13	2.4	4.01	3.57	1.49

Table 5: Absolute percentage error of the expected percentage of urgent and non-urgent patients exceeding n weeks of access delay.

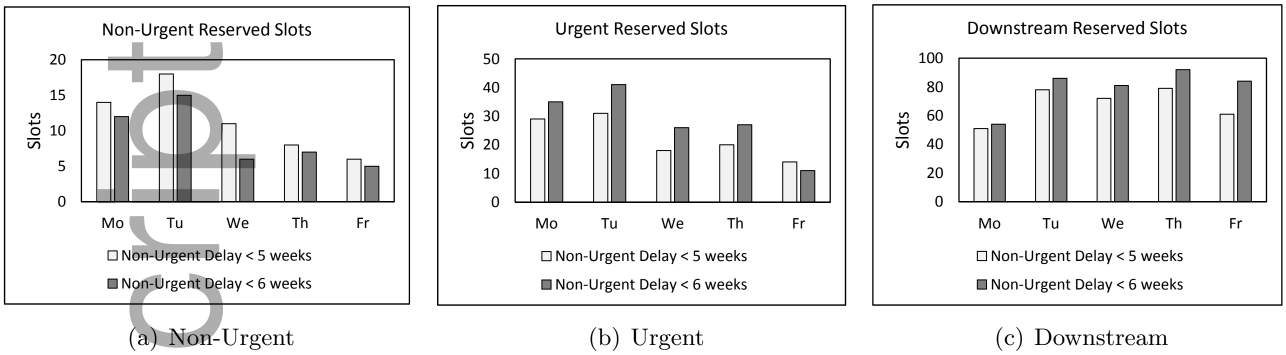
interest to our industry partner. This scenario is motivated by the Fig. 1, where urgent patients were experiencing longer access delays than non-urgent patients. The key information is reproduced for the particular service we studied in Fig. 9(a).

5.3.1 Minimizing Mean Access Delay

We began solving this problem by consulting our industry partners to obtain bounds on the various competing metrics modeled in APT. From these discussions, we selected APT settings that minimize mean access delay for urgent patients under the following constraints: (1) mean access delay for non-urgent cases should not exceed 5 weeks (25 workdays since we do not count weekends), (2) overtime is used on less than 10% of workdays, and (3) the expected number of appointments performed in overtime is less than 5 per day. The resulting template is shown in Fig. 8(a). Running our simulation using this template demonstrates a 26.9% decrease in mean access delay for urgent patients from 24.59 to 17.98 days (from close to 5 weeks down to 3.5 weeks on average). The trade-off of this improvement was an increase of only 1.6 days (7.3%) in mean access delay for non-urgent patients from 23.21 to 24.92 days. The probability of overtime under this schedule was 9.75%. Prior improvement projects conducted by the partner organization were not able to achieve the level of benefit offered by our new template design.

In this analysis we minimized the mean access delay for urgent patients while constraining the mean access delay for non-urgent patients. An alternative approach could minimize a weighted average of mean access delays for both urgent and non-urgent patients, while keeping the other constraints the same. For instance, if we choose the weights to be 73% on urgent and 27% on non-urgent mean access delay, which matches the fractions of the patient classes arriving to Neurology for root appointments (placing more weight on urgent patients), then the results are instead a 16.1% decrease in urgent mean access delay (from 24.59 days to 20.62 days) and a 5.8% decrease is non-urgent mean access delay (from 23.21 days to 21.86 days), while the probability of overtime of 9.87% is very close to the original formulation. In practice, management preferred the constrained approach, because they are able to set service-level targets which are more interpretable and easier work with than objective function weights. For the weighted objective instance reported here, the weights were somewhat arbitrarily chosen. Rather than trying to tune the weights, management was more comfortable working with constraints, which more directly reflect management's strategic goals. All future results use the constrained approach.

Another item of keen interest for our industry partner was to understand how the templates change when further relaxing the constraints on non-urgent patient access delay. For this what-if analysis, we increased the limit on mean access delay for non-urgent patients from 5 weeks to 6 weeks (Fig. 8).



(a) Non-Urgent

(b) Urgent

(c) Downstream

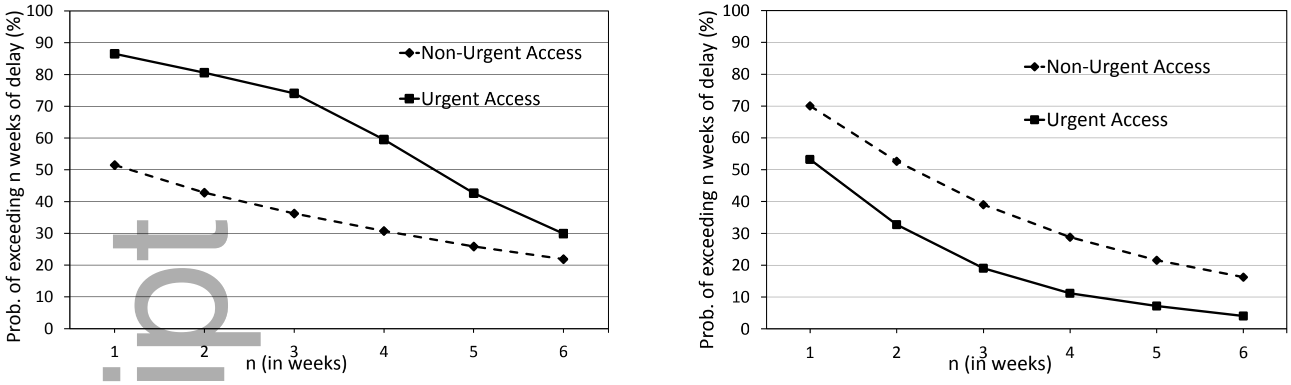
Figure 8: Optimized Neuro template (in number of appointments slots) under the constraints defined in Sec. 5.3.1 when the mean access delay constraint on non-urgent patients is increased from (Light) 5 to (Dark) 6 weeks.

To provide better urgent patient access, the new template reduces the overall slot reservations for non-urgent patients. Of particular interest to our partner organization, is that, in addition to reserving more urgent slots, the template also reserves more capacity for downstream appointments. This is because urgent patients tend to use more downstream resources than non-urgent cases due to their unknown condition and/or case complexity. Thus, the increase in urgent slots earlier in the week results in the need for more downstream appointments towards the middle and end of the week. These subtleties are easily overlooked without the aid of a decision support tool such as APT.

We conclude this analysis and segue to the next section with a more detailed examination of *how* the optimization was able to achieve this reduction in mean access delay. Specifically, optimizing mean access delay may leave some patients with very long delays while others have unnecessarily short delays. Thus, it is important to also examine service level metrics. Fig. 9 shows the complementary c.d.f. of access delays for urgent and non-urgent patients before (Fig. 9(a)) and after (Fig. 9(b)) optimization. This figure shows that, by focusing on the mean access delay, the optimization still leaves more than 20% of non-urgent patients with delays longer than 5 weeks and almost 20% of urgent patients with delays longer than 3 weeks, which was considered undesirable by our partner organization. Our partners showed great interest in ensuring that long waits are avoided in most cases (i.e., they are interested in high service levels). Hence, in the next section, we examine service level type constraints and how APT can control not just the mean access delays, but also the distribution of delays.

5.3.2 Controlling the Distribution of Access Delay

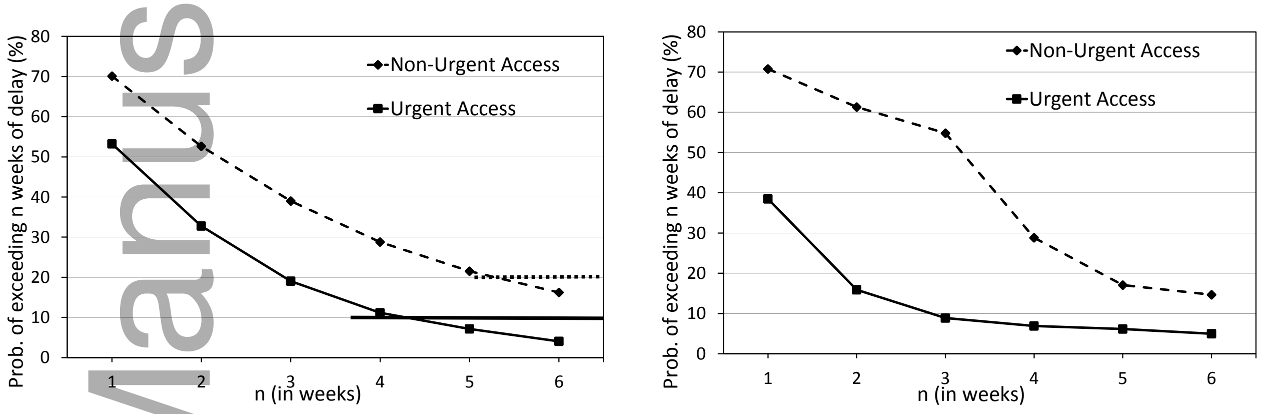
After seeing the results of the first template (see Fig. 9), the managers of that service indicated that, even though their initial constraints were met, it was unacceptable to have a service level in which more



(a) Historical access delay

(b) Optimized access delay

Figure 9: Results of a case study of the Neurology service demonstrating the ability to control the distribution (complementary c.d.f.) differentiating targets for urgent and non-urgent patients.



(a) Historical access delay

(b) Optimized access delay

Figure 10: Impact on the access delay violation probability (service level) curves for Neuro when two constraints (horizon lines in the left part of the figure) are imposed on the system: (Left) access delay curves under the constraints defined in Sec. 5.3, and (Right) access delay curves when we wish that (i) no more than 10% of the urgent patients wait longer than three weeks to get a root appointment and (ii) no more than 20% of the non-urgent patients wait longer than five weeks to get a root appointment.

than 10% of their urgent patients would have to wait longer than three weeks to get a root appointment. However, improving urgent patient service levels has a trade-off – negatively impacting the non-urgent access delay. Hence, our partners also requested that we add a service level constraint assuring that no more than 20% of their non-urgent cases wait longer than 5 weeks.

To incorporate these additional requirements, we add the following additional constraints to the previous model: (4) no more than 10% of the urgent cases will exceed three weeks access delay for a root appointment (represented by the horizontal solid line in Fig. 10(a)) and (5) no more than 20% the non-urgent cases will exceed five weeks access delay for a root appointment (represented by the horizontal dotted line in Fig. 10(a)). Fig. 10(b) shows the complementary c.d.f. of access delays resulting from the new optimization.

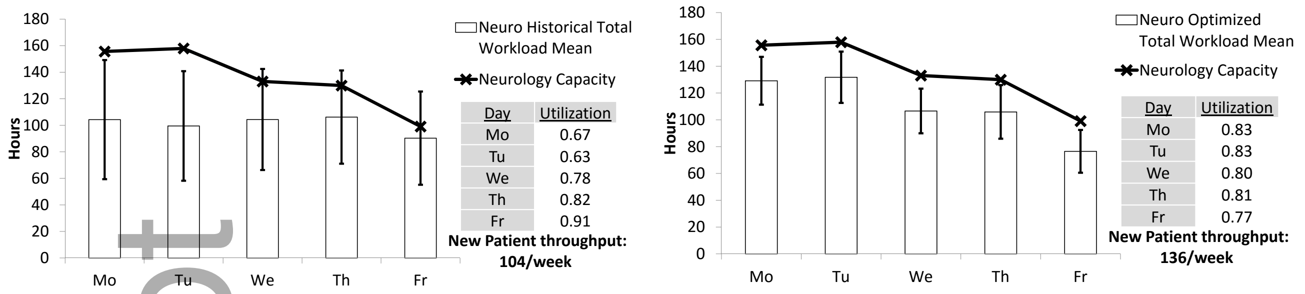
For urgent patients, note that now only 40% must wait longer than a week for a root appointment (as opposed to 55% under the original optimization) and the curve drops more sharply in the first 3 weeks to ensure that at most 10% of urgent patients wait longer than 3 weeks. This makes it more difficult for a non-urgent patient to obtain a root appointment in under three weeks: the number of non-urgent cases that will wait longer than three weeks for a root appointment increases from 40% to 54%. However, by including the service level target on non-urgent patients as well, there is a noticeable drop in the non-urgent access delay curve after three weeks to ensure that fewer than 20% of non-urgent patients wait longer than 5 weeks. By incorporating the ability to control probability distributions on access delay, APT provides a far more precise tool for managing customer service requirements in such complex service systems.

5.4 Increasing Urgent Patient Throughput

Another motivating factor for pursuing this research agenda was a request from both GIM and Neurology to increase the volume of urgent patients. In this section, we increase the volume of urgent root appointment requests while generating templates according to the optimization presented in Sec. 5.3.1. The optimization increases throughput as much as possible without worsening access delay or overtime as compared to historic levels. Further, we show that, by varying constraints on access delay and overtime, APT is able to provide managers with richer decision support in the form of efficient frontiers. This supports managerial decisions surrounding how much and what class (e.g., urgent vs non-urgent) of access they are willing to sacrifice to increase patient throughput.

For Neurology, urgent patient throughput was maximized under the following constraints on non-urgent patients: (1) no more than 30% of patients exceed 1 week access delay, (2) no more than 20% exceed 2 weeks access delay, and (3) no more than 10% exceed three weeks access delay. The results of this study are summarized in Fig. 11, which displays both the historical total workload in Neurology (Fig. 11(a)) and the optimized workload with increased throughput (Fig. 11(b)). The optimization results in 31% increase in throughput relative to the historical rate. At the same time, expected utilization is higher, and the standard deviation of total workload is lower. This is despite the fact that urgent patient root appointments on average result in the highest downstream workload requirements. APT is capable of producing a more controlled plan that increases expected utilization in the clinic by reducing workload variability and also by smoothing workloads relative to average capacity (see Fig. 11(b)).

Using the method described above, we also created efficient frontiers for the GIM service, displayed in Fig. 12. For this study, we constrain access delays such that no more than 5% of urgent patient requests will wait more than 4 weeks for a root appointment, and all patient types will have access



(a) Historical

(b) Optimized

Figure 11: Comparison of workload mean and standard deviation relative to average capacity for (a) historical and (b) optimized throughput scenarios of 104 and 136 urgent patient consults per week, respectively. The whiskers represent ± one standard deviation.

delays no longer than 6 weeks. We also include constraints similar to those considered previously on total overtime. In Fig. 12, we consider the impact of increasing urgent patient volume (vertical axis) subject to service level guarantees on access delay for non-urgent patients (indicated by the different curves) and constraints on the probability of overtime (along the horizontal axis). In the figure, we denote the non-urgent patient access delay metric by the term "Service Level x%," by which we mean that no more than x% of non-urgent patients will have access delays longer than 4 weeks.

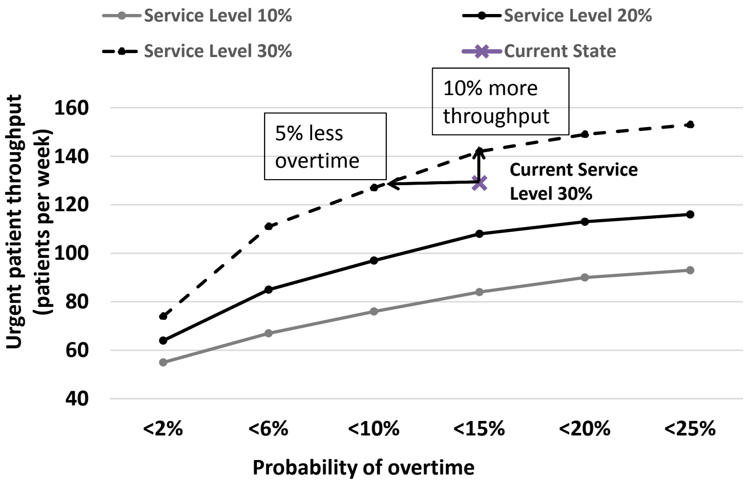


Figure 12: Efficient frontier comparing probability of overtime (horizontal axis), urgent patient throughput (vertical axis), and non-urgent patient access delays of more than 4 weeks – service level x% means that less than x% of non-urgent patients waited longer than 4 weeks (three curves).

These frontiers accurately capture the inherent trade-offs in key patient mix decisions. For example, with 10% overtime probability, the optimal schedule can achieve an urgent patient throughput of either (1) ∼80 urgent patients per week while ensuring no more than 10% of non-urgent patients wait more than 4 weeks to get an appointment, (2) ∼100 urgent patients per week if we allow up to 20% of non-urgent patients to wait more than 4 weeks, or (3) ∼130 urgent patients per week if we allow up to 30% of non-urgent patients to wait more than 4 weeks for their root appointment. In current practice (denoted by the '×' in Fig. 12) 30% of non-urgent patients wait more than 4 weeks, the overtime probability is at least 15% (a lower bound on current average overtime provided by our industry partners), and throughput 129 patients per week. This can be improved in a number of ways. For example, Fig. 12 shows that it would be possible to increase throughput by 10% with the same level of overtime and non-urgent access delay, or decrease overtime probability by 5% (absolute) with the same level of throughput and access delay. There are other ways to improve upon the current state as well by jumping to other curves. These frontiers are efficient solutions that provide a rich decision framework regarding the effects of changing patient mix for strategic decisions that have traditionally been made in an ad-hoc, trial and error manner.

Effective sensitivity analysis is a critical part of meeting the advisory and managerial decision support goals. Based on our interactions with physicians and the managerial staff, it seems that this sensitivity analysis feature is a key component of this new methodology. Analyzing many different template scenarios allows management to incorporate their experience into a trade-off analysis, giving them the control and information needed to make effective decisions.

5.5 The Value of the Integrated Solution

For the purpose of comprehensive and rapid diagnosis and treatment plan design, patients are often scheduled for a root appointment in a diagnostic department. Based on the results of the initial tests, new appointments are generated at other departments for deeper diagnosis and/or to begin treatment design. In this section, we demonstrate the value of the integrated solution by comparing it with a model that optimizes services independently, which we call the siloed solution. This comparison is a conservative estimate of the benefit of APT, since the siloed solution presented here is still the result of an optimization.

We first solve the optimization presented in Section 5.3.1 to minimize mean access delay in Neurology independent of the other services. We then compute the difference in access delay and overtime for all three services (Neuro, GI, and GIM), comparing the integrated and siloed solutions. Table 6 shows the performance metrics of the siloed solution subtracted from the integrated. Negative values indicate an improvement of the integrated solution over the siloed solution. The arrivals per week, which are the same for both the integrated and the siloed scenarios, are also provided.

Table 6 demonstrates the problems associated with independent management of integrated services discussed in the introduction. As compared to the integrated solution, the siloed solution has the strongest negative impact on GIM. For GIM, the integrated solution decreases the probability of overtime by 5.51%, decreases the mean access delay for urgent patients by 7%, and decreases the mean access

Table 6: Absolute (percent) difference for mean access delay in days and absolute difference for probability of overtime of the integrated solution relative to the siloed solution. (Prob. is an absolute value, multiplied by 100.) Negative values indicate an improvement.

	Arrivals/Week		Diff. Mean Access Delay in Days (%)		Diff. Prob. Overtime
	Urgent	Non-Urgent	Urgent	Non-Urgent	
Neuro	104	39	3.36 (18.69%)	0.03 (0.12%)	2.23
GI	53	132	-0.05 (-0.41%)	0.60 (2.46%)	-0.20
GIM	130	85	-2.63 (-7.24%)	-4.67 (-19.9%)	-5.51

delay for non-urgent patients by 20%. GI experiences little impact because few Neurology patients have downstream appointments in GI. The mean access delay for Neurology is, of course, smaller for the siloed solution since it ignores other departments. From a system's perspective, the integrated solution is better. In aggregate, across the three departments there is an average reduction of 0.6 days (2.3% decrease) of mean access delay per patient (calculated by multiplying the change in mean access delay by number of patients in each category in Table 6). Other benefits include a 1.7% reduction in the chance a patient will need to be served in overtime across the three departments.

Next, we consider a scenario where the departments have greater interdepartmental flows. Some clusters of services at our partner institution have greater connectivity of downstream appointments than those exhibited by Neurology, GI, and GIM; hence, we construct a counterfactual (that is, a hypothetical example) where we increase the probability that a patient with a root appointment in Neurology requires subsequent downstream appointments in GI and GIM. Historically, every root appointment in Neurology generated on average 0.13 and 0.03 downstream appointments in GIM and GI respectively. In our counterfactual study, we increase this to 0.26 and 0.15 downstream appointments in GIM and GI respectively. Using an integrated model as opposed to the siloed model improves the aggregate mean access delay (across all 3 departments) by 2.7 days (11% decrease), with mean access delays reduced by 3.0 days (13% decrease) for non-urgent patients, and 2.3 days (9.0% decrease) for urgent patients. The overtime probability is reduced by 5.0%. When departments are more connected, the integrated solution demonstrates even greater gains across all aggregate metrics. Table 7 summarizes the results for each department in the same format as Table 6.

Table 7: Counterfactual case absolute (percent) difference of the integrated relative to the siloed solution. (Prob. is an absolute value, multiplied by 100.)

	Arrivals/Week		Diff. Mean Access Delay in Days (%)		Diff. Prob. Overtime
	Urgent	Non-Urgent	Urgent	Non-Urgent	
Neuro	104	39	7.07 (37.0%)	2.68 (12.7%)	7.26
GI	53	132	-4.90 (-37.6%)	-5.20 (-21.5%)	-4.02
GIM	130	85	-8.77 (-24.2%)	-2.30 (-9.8%)	-13.93

For this study, computations were performed using IBM CPLEX on a computer with an Intel Xeon E5-2640v3 2.6 GHz processor. Runtimes for the integrated solutions of the three departments ranged from 1.5 hours up to 12-15 hours in some cases. In contrast, siloed solutions ranged from 17 minutes to 2-2.5 hours. The wide range of run times depended on how tight the constraints were. For tighter constraints, the algorithm may spend 90% of the time trying to determine a feasible solution. Since this is a planning model, these runtimes are acceptable for practice.

6 Conclusion

This work contributes to the sparse research on advance capacity planning methods supporting effective control of access delay for appointments in integrated outpatient care delivery systems with multiple patient classes that have multi-visit stochastic itineraries in a network of specialist services. We take a novel approach that linearly approximates system congestion to enable tractable optimization of capacity planning appointment templates to control access delays via mean and service level constraints. Our model can control not only mean delay, but even the shape the distribution of wait times, which allows for much finer control over the delays experienced by each class of patient.

To solve this complex stochastic optimization, we transform the model into a deterministic mixed integer program, which allows for tractable optimization and the ability to model many performance constraints required in practice. This new approach promises to increase the ability to manage complex tradeoffs involving (1) operational efficiency, (2) access delays for urgent patients, and (3) the amount of network overtime. These objectives are not new to leading organizations; however, advanced methods to achieve metric targets were not previously available. The alternative to an optimization approach is to employ intuition driven policies that are evaluated by simulation. However, it is extremely difficult to obtain well-performing solutions due to the size and complexity of the policy space. While optimizing a network of three services was computationally feasible, large networks may suffer from tractability, which represents an area for future research.

To apply APT in practice, we suggest the following process: (1) identify patient types and priorities, (2) define metrics and target levels, (3) calculate resource requirements for each patient type, (4) identify connected bundles of services, (5) generate templates, workload forecasts, and trade-off curves, and (6) evaluate and approve the final templates. In the first step, management must determine a set of patient characteristics and a priority ordering that meets the clinical and business objectives of the organization. These characteristics for a class of patients may include: condition, complexity, severity, convenience (e.g., distance traveled to receive service), and physician research, and practice goals.

Acknowledgements

We thank Dr. Thomas Rohleder, Mark Hayward, and Brian Crum, M.D., of Mayo Clinic for their leadership that advanced this research project. The work of Mark P. Van Oyen and Jivan Deglise-Hawkinson was supported in part by the grants NSF CMMI-1233095 and 1548201 and NIH UL1TR000433. NSF and NIH had no role in the design, conduct, or writing of this research.

References

Adan, I., J. Bekkers, N. Dellaert, J. Vissers, X. Yu. 2009. Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Mgt. Sci.* **12**(2) 129–141.

Akkan, C. 1997. Finite-capacity scheduling-based planning for revenue-based capacity management. *Eur. J. Oper. Res.* **100**(1) 170–179.

American Hospital Association. 2015. Connecting the dots along the care continuum. White paper, `http://www.aha.org/content/15/15carecontinuum.pdf`.

Bekker, R., P.M. Koeleman. 2011. Scheduling admissions and reducing variability in bed demand. *Health Care Mgt. Sci.* **14**(3) 237.

Cayirli, T., E. Veral. 2003. Outpatient scheduling in health care: A review of literature. *Prod. Oper. Manag.* **12**(4) 519–549.

Chow, V.S., M.L. Puterman, N. Salehirad, W. Huang, D. Atkins. 2011. Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Prod. Oper. Manag.* **20**(3) 418–430.

Deglise-Favre-Hawkinson, J. 2015. Access and resource management for clinical care and clinical research in multi-class stochastic queueing networks. Ph.D. thesis, University of Michigan.

Erdelyi, A., H. Topaloglu. 2010. Approximate dynamic programming for dynamic capacity allocation with multiple priority levels. *I.I.E. Trans.* **43**(2) 129–142.

Feldman, J., N. Liu, H. Topaloglu, S. Ziya. 2014. Appointment scheduling under patient preference and no-show behavior. *Oper. Res.* **62**(4) 794–811.

Gerchak, Y., D. Gupta, M. Henig. 1996. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Manage. Sci.* **42**(3) 321–334.

Gocgun, Y., A. Ghate. 2012. Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Comput. Oper. Res.* **39**(10) 2323–2336.

Gocgun, Y., M.L. Puterman. 2014. Dynamic scheduling with due dates and time windows: an application to chemotherapy patient appointment booking. *Health Care Mgt. Sci.* **17**(1) 60–76.

Gupta, D., B. Denton. 2008. Appointment scheduling in health care: Challenges and opportunities. *I.I.E. Trans.* **40**(9) 800–819.

Gupta, D., L. Wang Liu. 2008. Revenue management for a primary-care clinic in the presence of patient choice. *Oper. Res.* **56**(3) 576–592.

Gupta, D., M.K. Natarajan, A. Gafni, L. Wang, D. Shilton, D. Holder, S. Yusuf. 2007. Capacity planning for cardiac catheterization: a case study. *Health Policy* **82**(1) 1–11.

Helm, J.E., M.P. Van Oyen. 2014. Design and optimization methods for elective hospital admissions. *Oper. Res.* **62**(6) 1265–1282.

Helm, J.E., M.P. Van Oyen. 2016. Capacity optimization to improve itinerary completion in a destination medical center. Working paper.

Herbots, J., W. Herroelen, R. Leus. 2010. Single-pass and approximate dynamic-programming algorithms for order acceptance and capacity planning. *J. Heuristics* **16**(2) 189–209.

Hsu, H-M., W-P. Wang. 2001. Possibilistic programming in production planning of assemble-to-order environments. *Fuzzy sets and Systems* **119**(1) 59–70.

Hulshof, P.J.H., R.J. Boucherie, E.W. Hans, J.L. Hurink. 2013. Tactical resource allocation and elective patient admission planning in care processes. *Health Care Mgt. Sci.* **16**(2) 152–166.

Hulshof, P.J.H., N. Kortbeek, R.J. Boucherie, E.W. Hans, P.J.M. Bakker. 2012. Taxonomic classification of planning decisions in health care: A structured review of the state of the art in OR/MS. *Health Systems* **1**(2) 129–175.

Kocher, R., N.R. Sahni. 2010. Physicians versus hospitals as leaders of accountable care organizations. *New. Engl. J. Med.* **363**(27) 2579–2582.

Lamiri, M., X. Xie, A. Dolgui, F. Grimaud. 2008. A stochastic model for operating room planning with elective and emergency demand for surgery. *Eur. J. Oper. Res.* **185**(3) 1026–1037.

Leung, K.K., W.A. Massey, W. Whitt. 1994. Traffic models for wireless communication networks. *IEEE J. Sel. Area Comm.* **12**(8) 1353–1364.

Mula, J., R. Poler, J.P. Garcia-Sabater, F.C. Lario. 2006. Models for production planning under uncertainty: A review. *Int. J. Prod. Econ.* **103**(1) 271–285.

Patrick, J., M.L. Puterman, M. Queyranne. 2008. Dynamic multi-priority patient scheduling for a diagnostic resource. *Oper. Res.* **56**(6) 1507–1525.

Saghafian, S., W.J. Hopp, M.P. Van Oyen, J.S. Desmond, S.L. Kronick. 2014. Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing Service Oper. Management* **16**(3) 329–345.

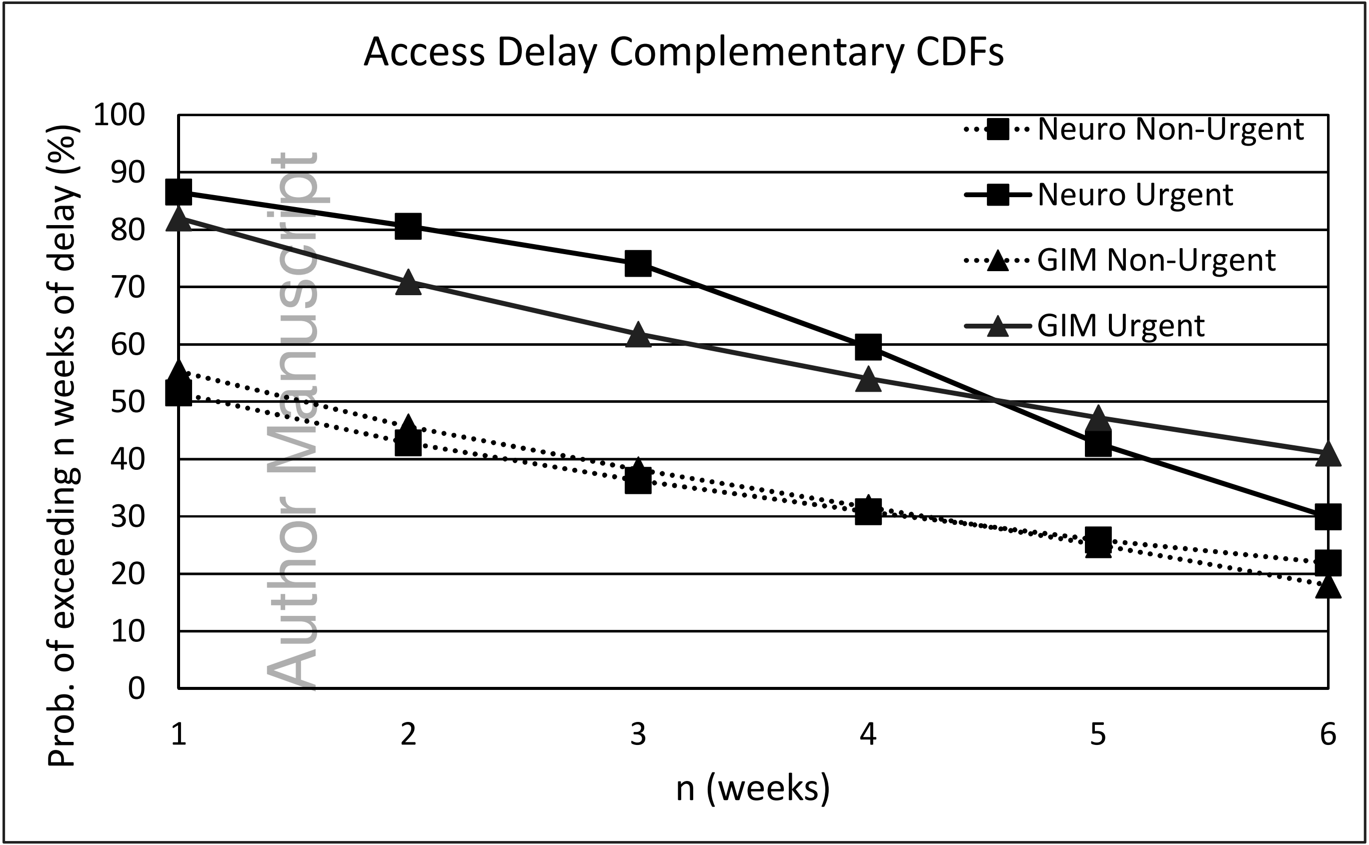
Sauré, A., J. Patrick, S. Tyldesley, M.L. Puterman. 2012. Dynamic multi-appointment patient scheduling for radiation therapy. *Eur. J. Oper. Res.* **223**(2) 573–584.

Talluri, K.T., G.J. Van Ryzin. 2006. *The Theory and Practice of Revenue Management*, vol. 68. Springer Science & Business Media.

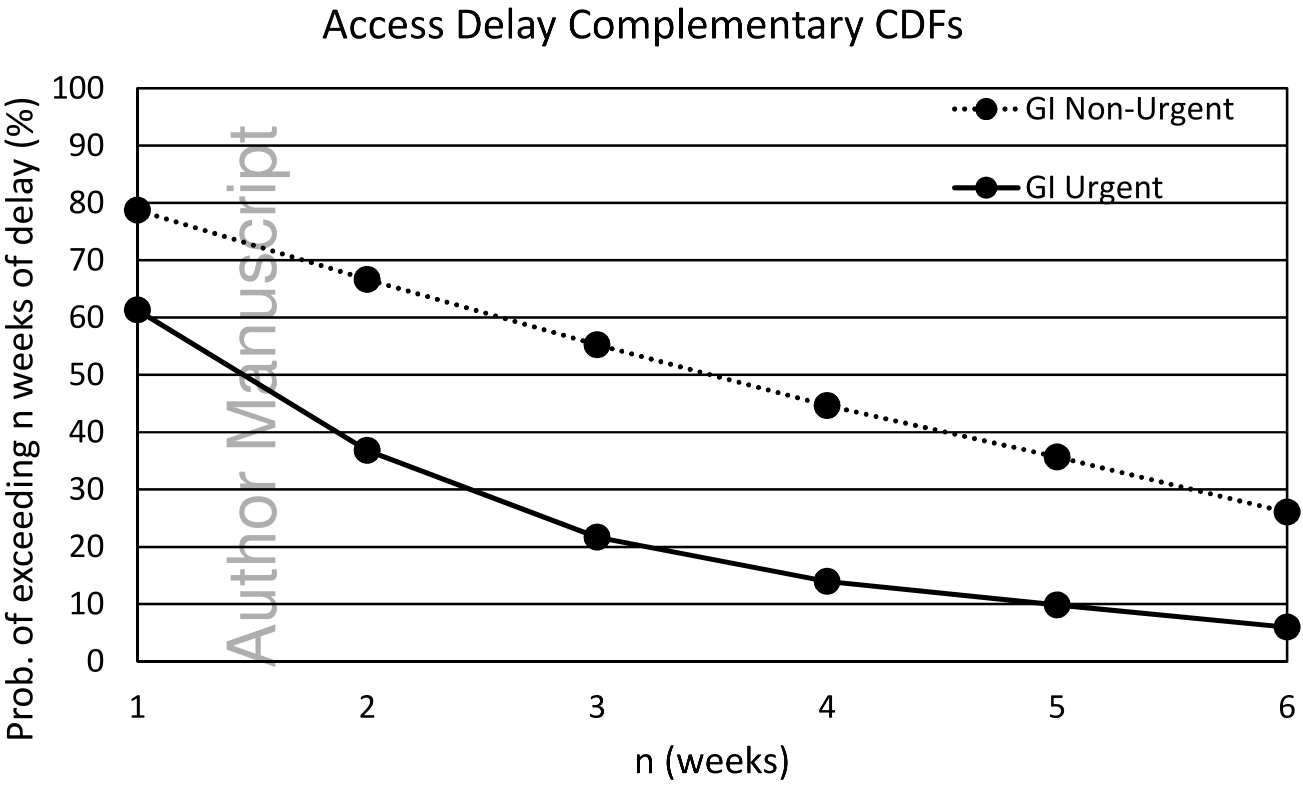
Truong, Van-Anh. 2015. Optimal advance scheduling. *Manage. Sci.* **61**(7) 1584–1597.

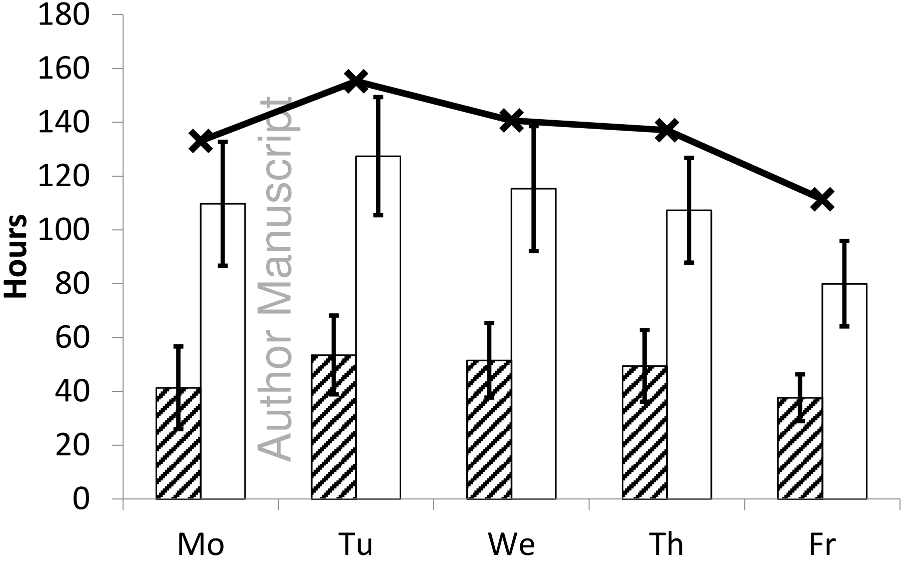
Turkcan, A., B. Zeng, M. Lawley. 2012. Chemotherapy operations planning and scheduling. *I.I.E. Trans. Healthc. Syst. Eng.* **2**(1) 31–49.

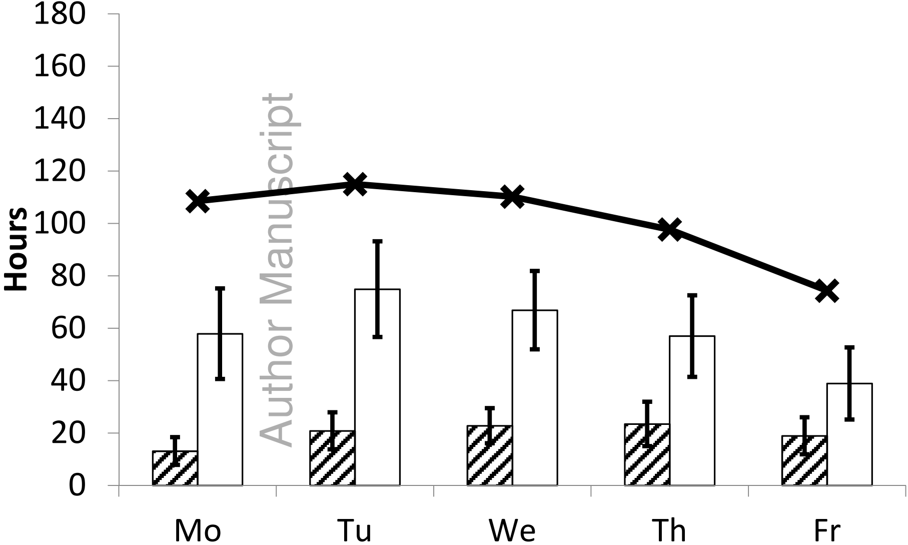
Vermeulen, I.B., S.M. Bohte, S.G. Elkhuizen, H. Lameris, P.J.M. Bakker, H.L. Poutré. 2009. Adaptive resource allocation for efficient patient scheduling. *Artif. Intell. Med.* **46**(1) 67–80.



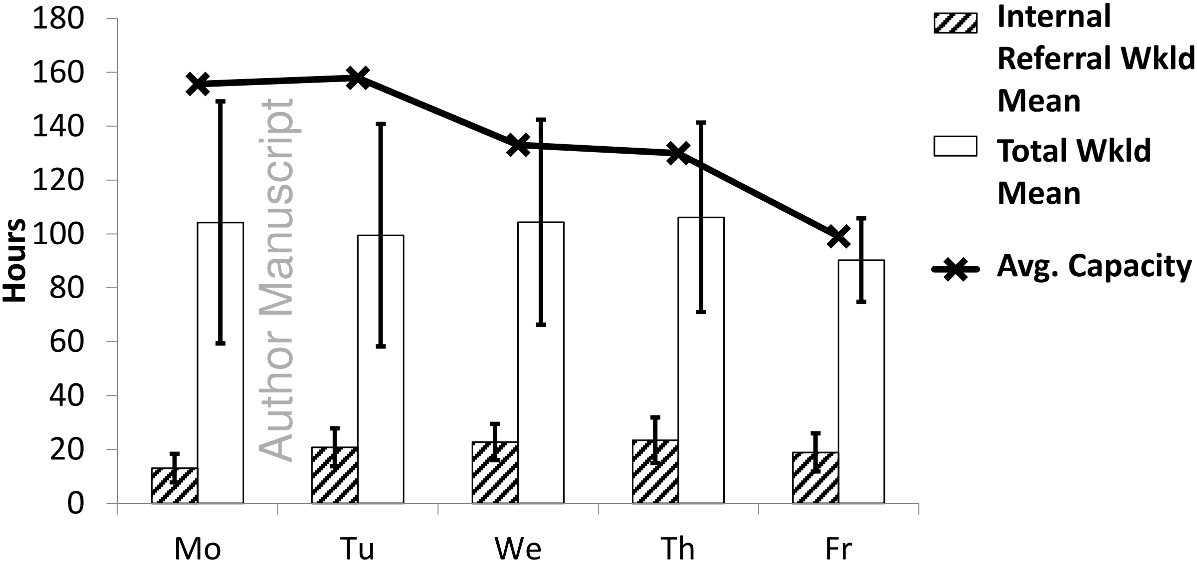
Access Delay Complementary CDFs
Prob. of exceeding n weeks of delay (%)
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
n (weeks)
Neuro Non-Urgent
Neuro Urgent
GIM Non-Urgent
GIM Urgent


Access Delay Complementary CDFs
Prob. of exceeding n weeks of delay (%)
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
n (weeks)
GI Non-Urgent
GI Urgent

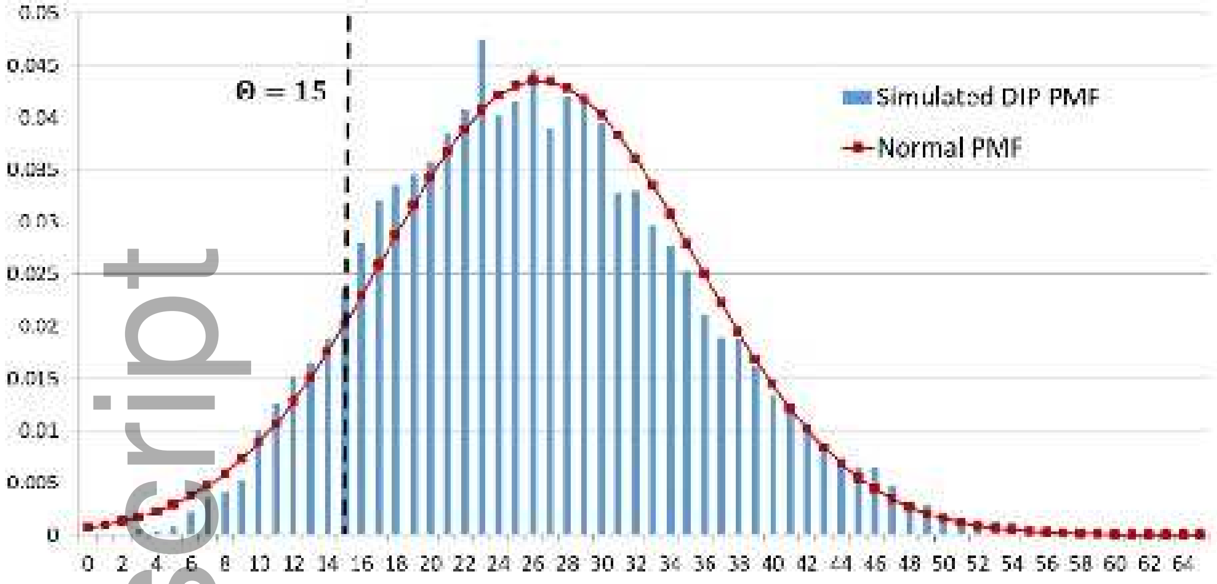

Hours
180
160
140
120
100
80
60
40
20
0
Mo
Tu
We
Th
Fr



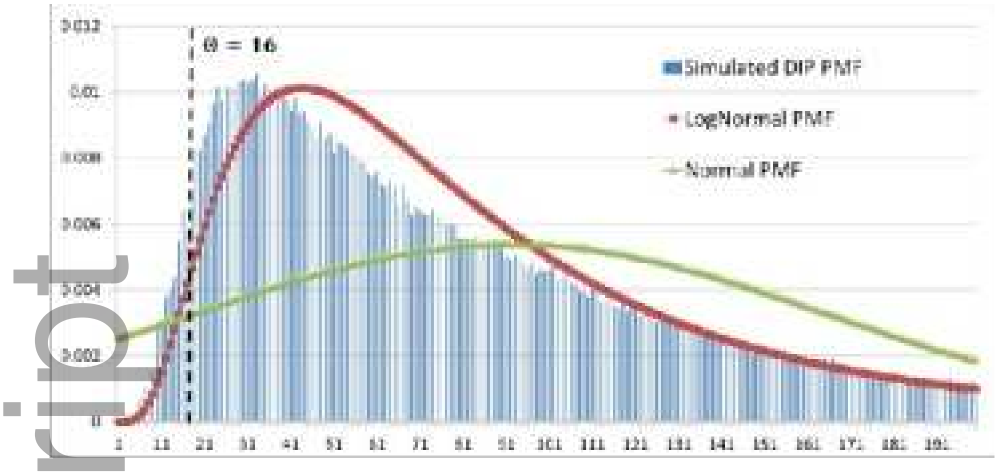
Hours
180
160
140
120
100
80
60
40
20
0
Mo
Tu
We
Th
Fr



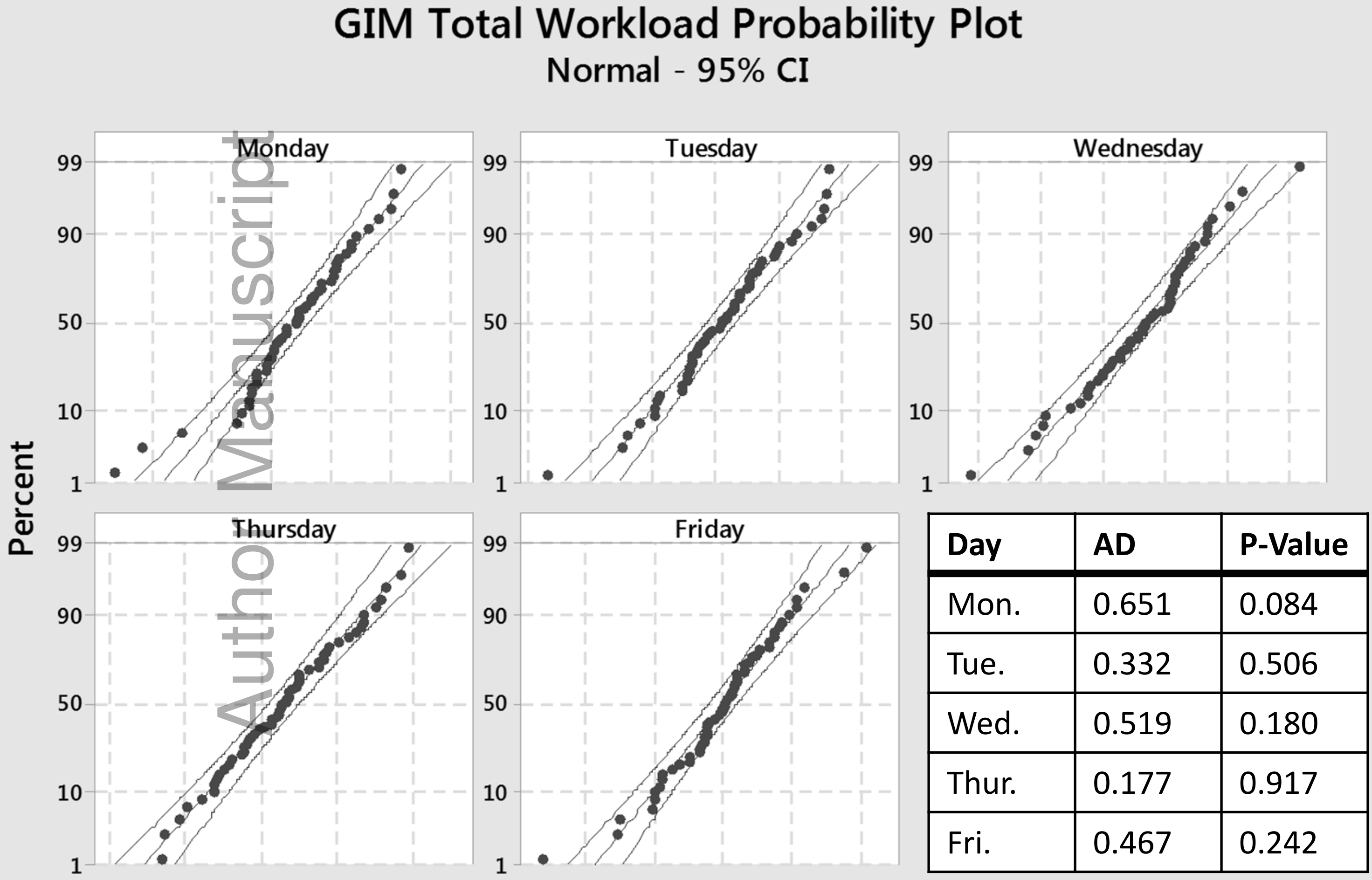
180
160
140
120
100
80
60
40
20
0
Hours
Mo
Tu
We
Th
Fr
▨ Internal Referral Wkld Mean
□ Total Wkld Mean
Avg. Capacity



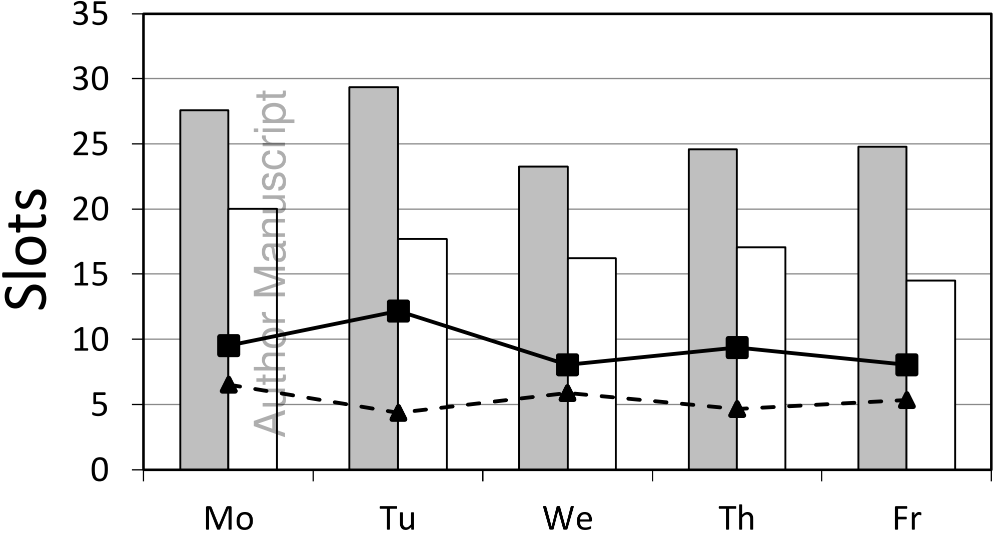
poms_12941_f3a.png

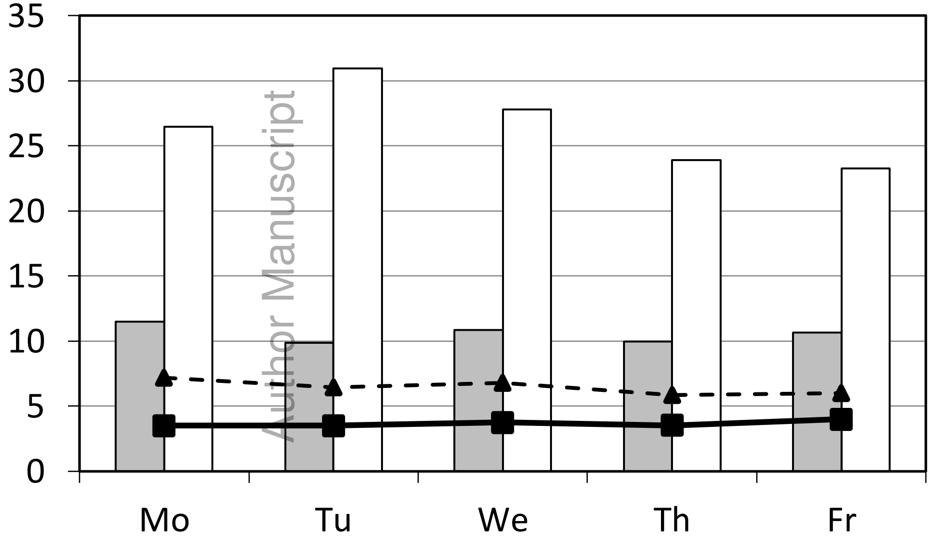


poms_12941_f3b.png

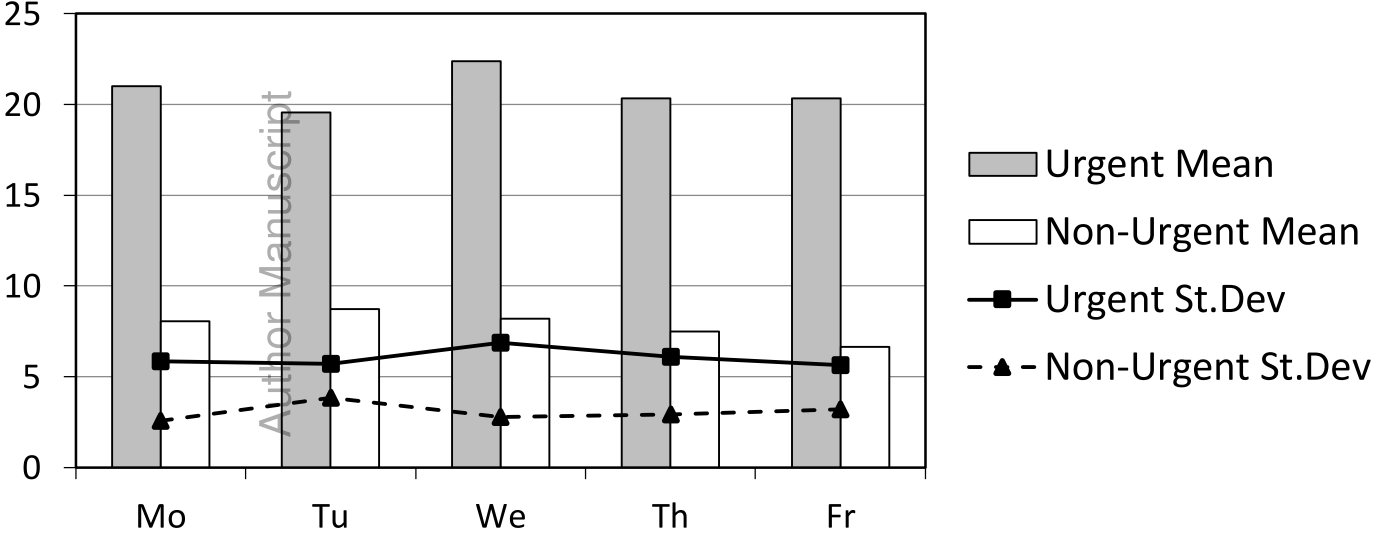


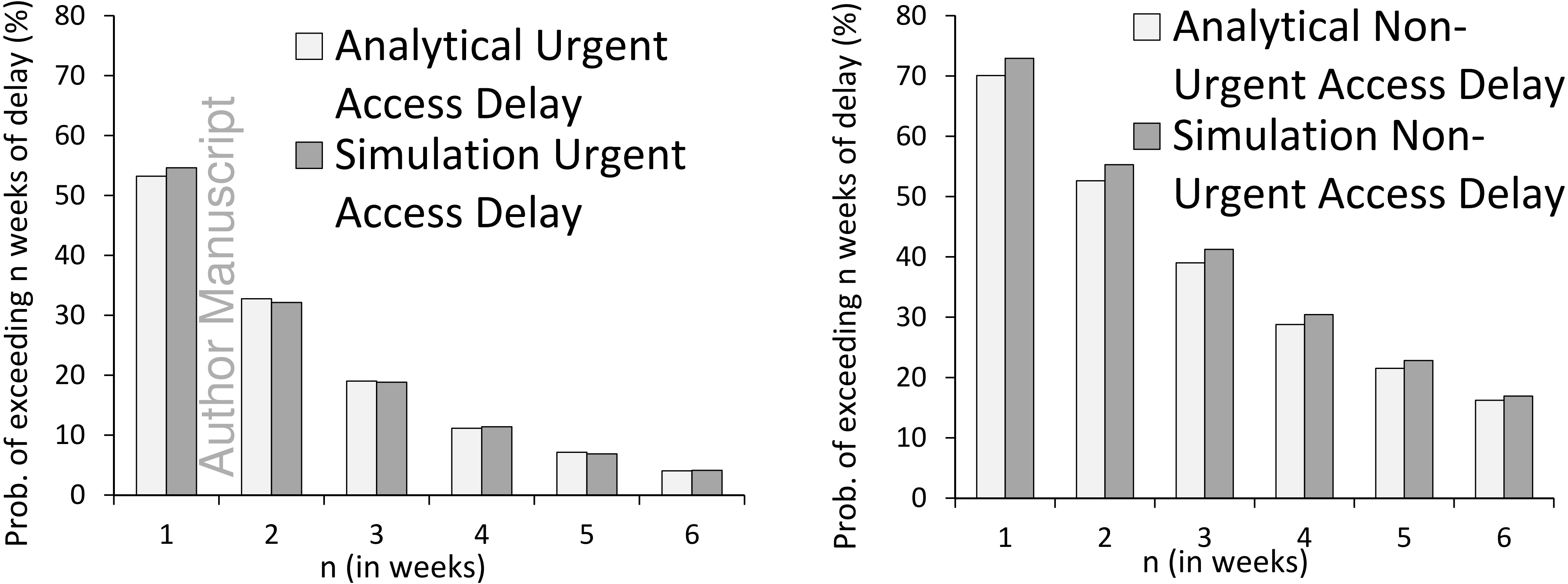
Day	AD	P-Value
Mon.	0.651	0.084
Tue.	0.332	0.506
Wed.	0.519	0.180
Thur.	0.177	0.917
Fri.	0.467	0.242

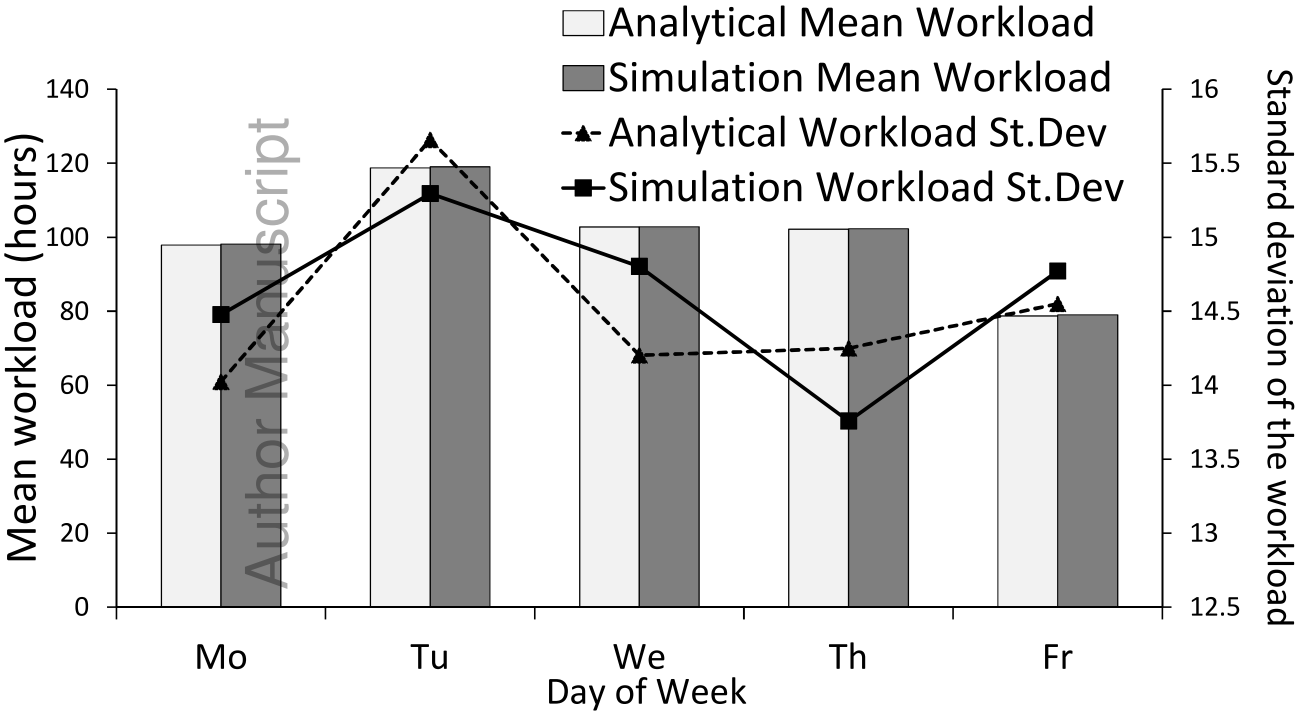

35
30
25
20
15
10
5
0
Slots
Mo
Tu
We
Th
Fr

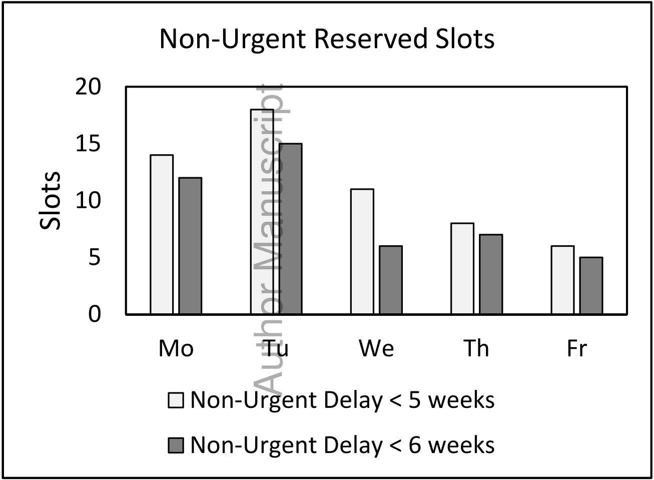


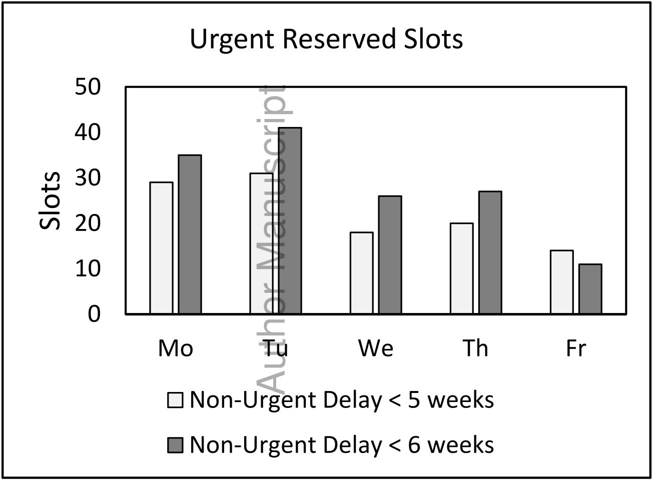
35
30
25
20
15
10
5
0
Mo
Tu
We
Th
Fr

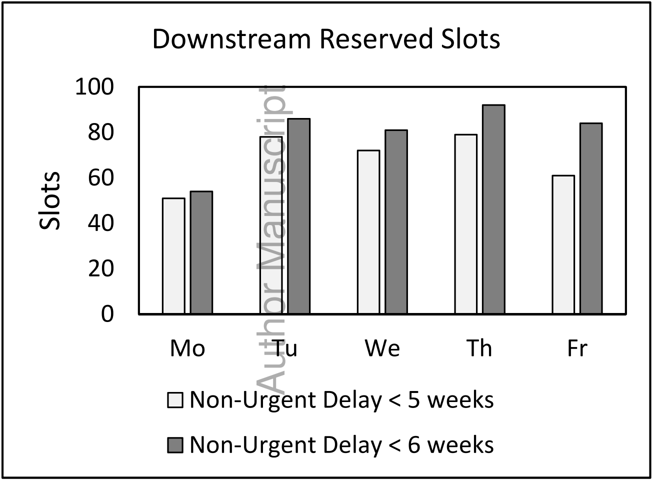

25
20
15
10
5
0
Mo
Tu
We
Th
Fr
Urgent Mean
Non-Urgent Mean
Urgent St.Dev
Non-Urgent St.Dev

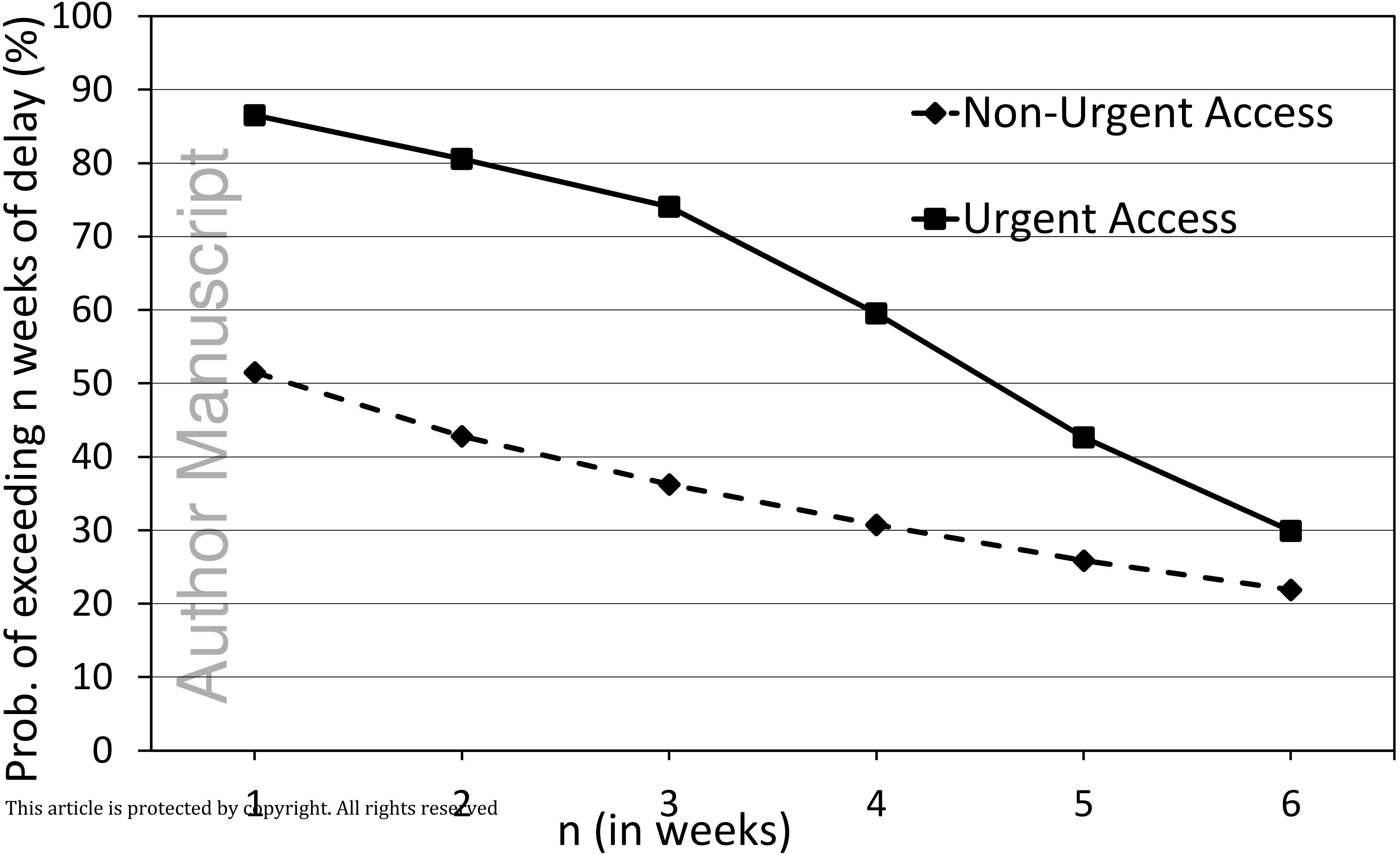

Prob. of exceeding n weeks of delay (%)
0
10
20
30
40
50
60
70
80
□ Analytical Urgent Access Delay
▨ Simulation Urgent Access Delay
1
2
3
4
5
6
n (in weeks)
Prob. of exceeding n weeks of delay (%)
0
10
20
30
40
50
60
70
80
□ Analytical Non-Urgent Access Delay
▨ Simulation Non-Urgent Access Delay
1
2
3
4
5
6
n (in weeks)

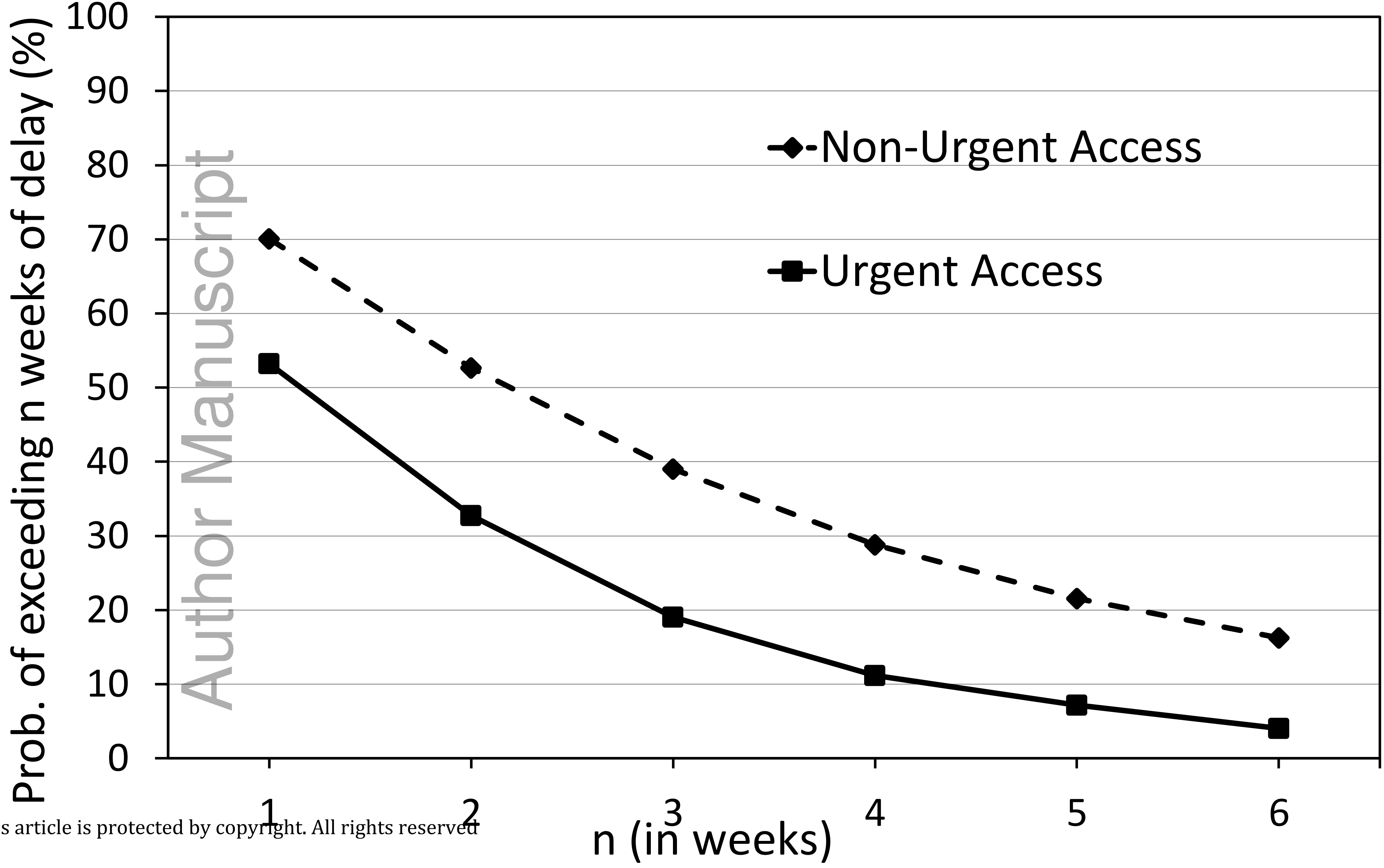

Analytical Mean Workload
Simulation Mean Workload
Analytical Workload St.Dev
Simulation Workload St.Dev
Mean workload (hours)
0
20
40
60
80
100
120
140
Standard deviation of the workload
12.5
13
13.5
14
14.5
15
15.5
16
Mo
Tu
We
Th
Fr
Day of Week

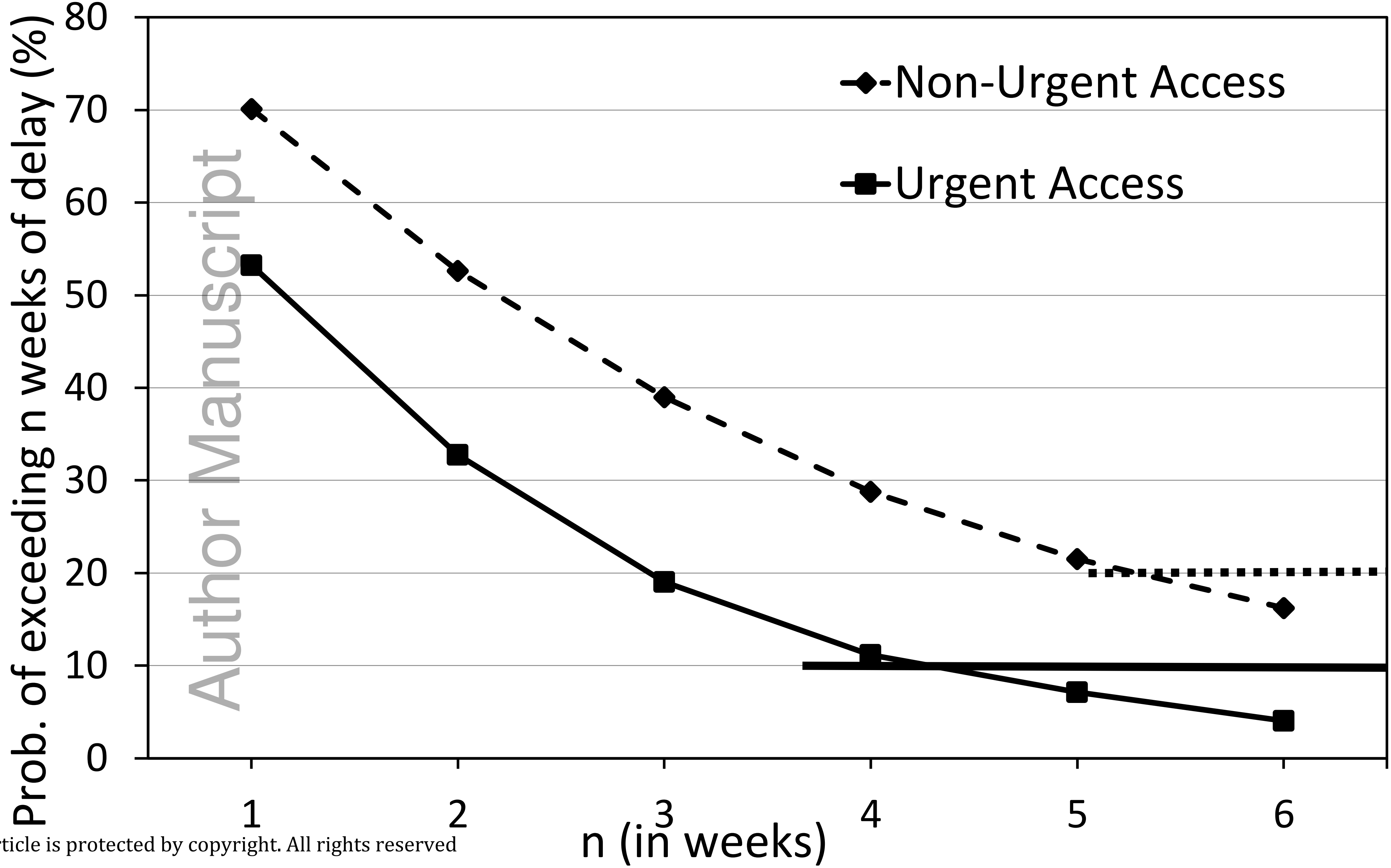

Non-Urgent Reserved Slots
Slots
20
15
10
5
0
Mo
Tu
We
Th
Fr
□ Non-Urgent Delay < 5 weeks
■ Non-Urgent Delay < 6 weeks

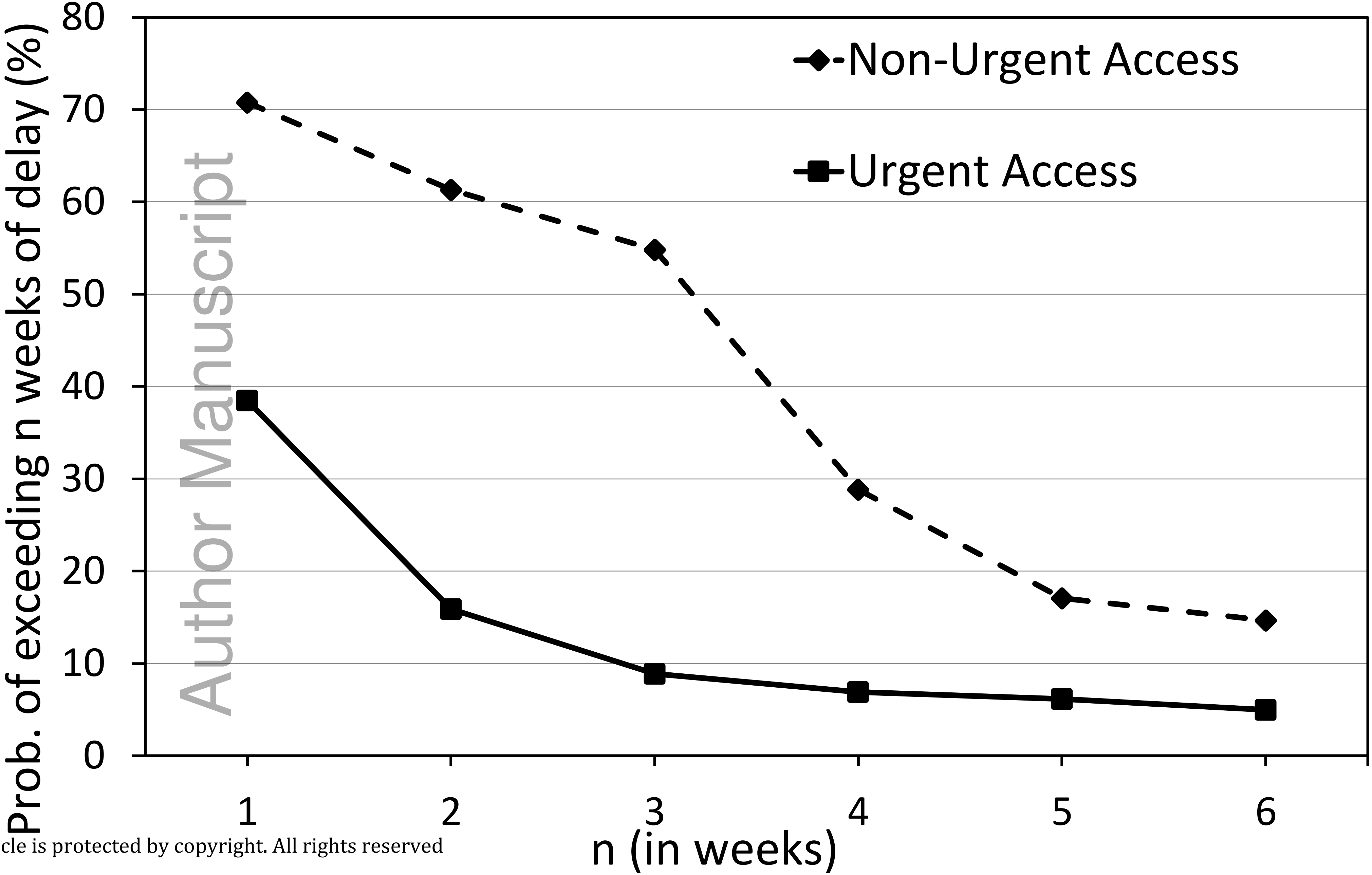

Urgent Reserved Slots
Slots
50
40
30
20
10
0
Mo
Tu
We
Th
Fr
□ Non-Urgent Delay < 5 weeks
■ Non-Urgent Delay < 6 weeks

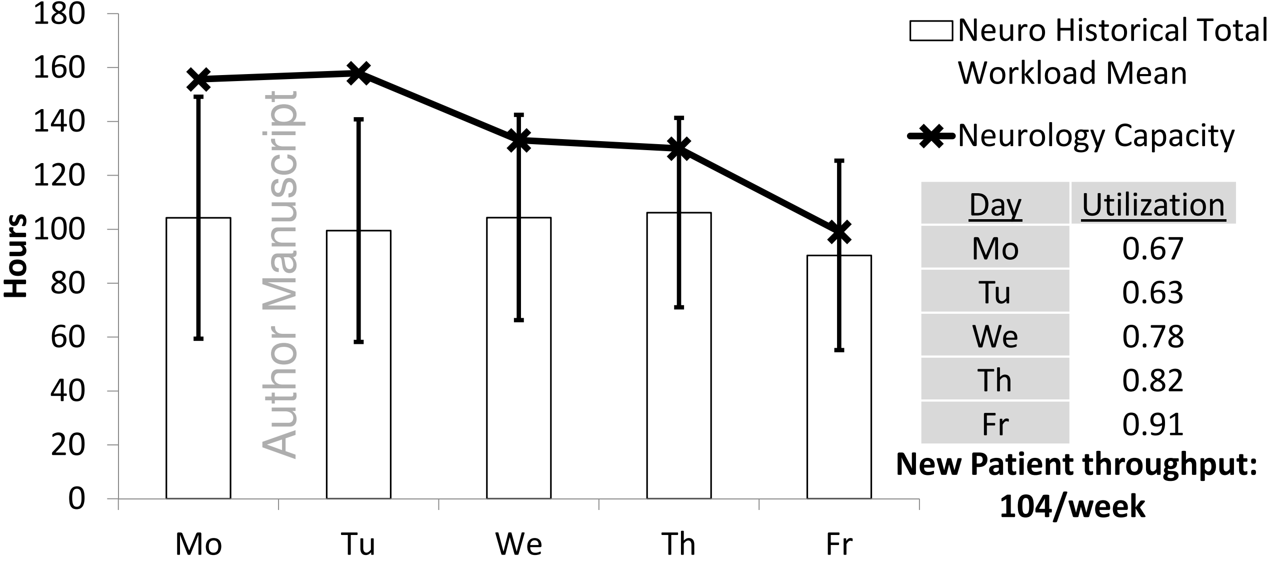

Downstream Reserved Slots
Slots
100
80
60
40
20
0
Mo
Tu
We
Th
Fr
□ Non-Urgent Delay < 5 weeks
■ Non-Urgent Delay < 6 weeks


100
90
80
70
60
50
40
30
20
10
0
Prob. of exceeding n weeks of delay (%)
1
2
3
4
5
6
n (in weeks)
Non-Urgent Access
Urgent Access


Prob. of exceeding n weeks of delay (%)
100
90
80
70
60
50
40
30
20
10
0
Non-Urgent Access
Urgent Access
1
2
3
4
5
6
n (in weeks)

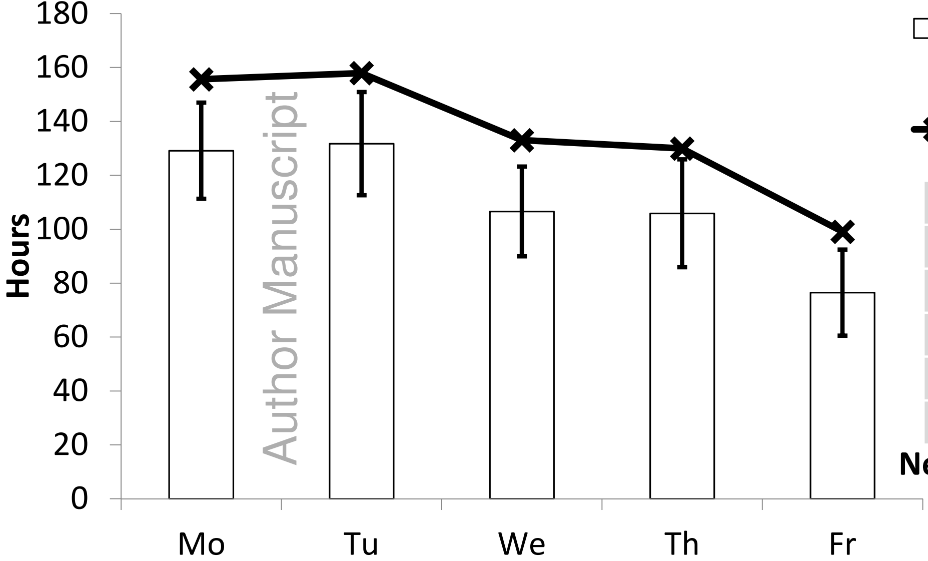

Prob. of exceeding n weeks of delay (%)
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
n (in weeks)
-Non-Urgent Access
Urgent Access


Non-Urgent Access
Urgent Access
Prob. of exceeding n weeks of delay (%)
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
n (in weeks)



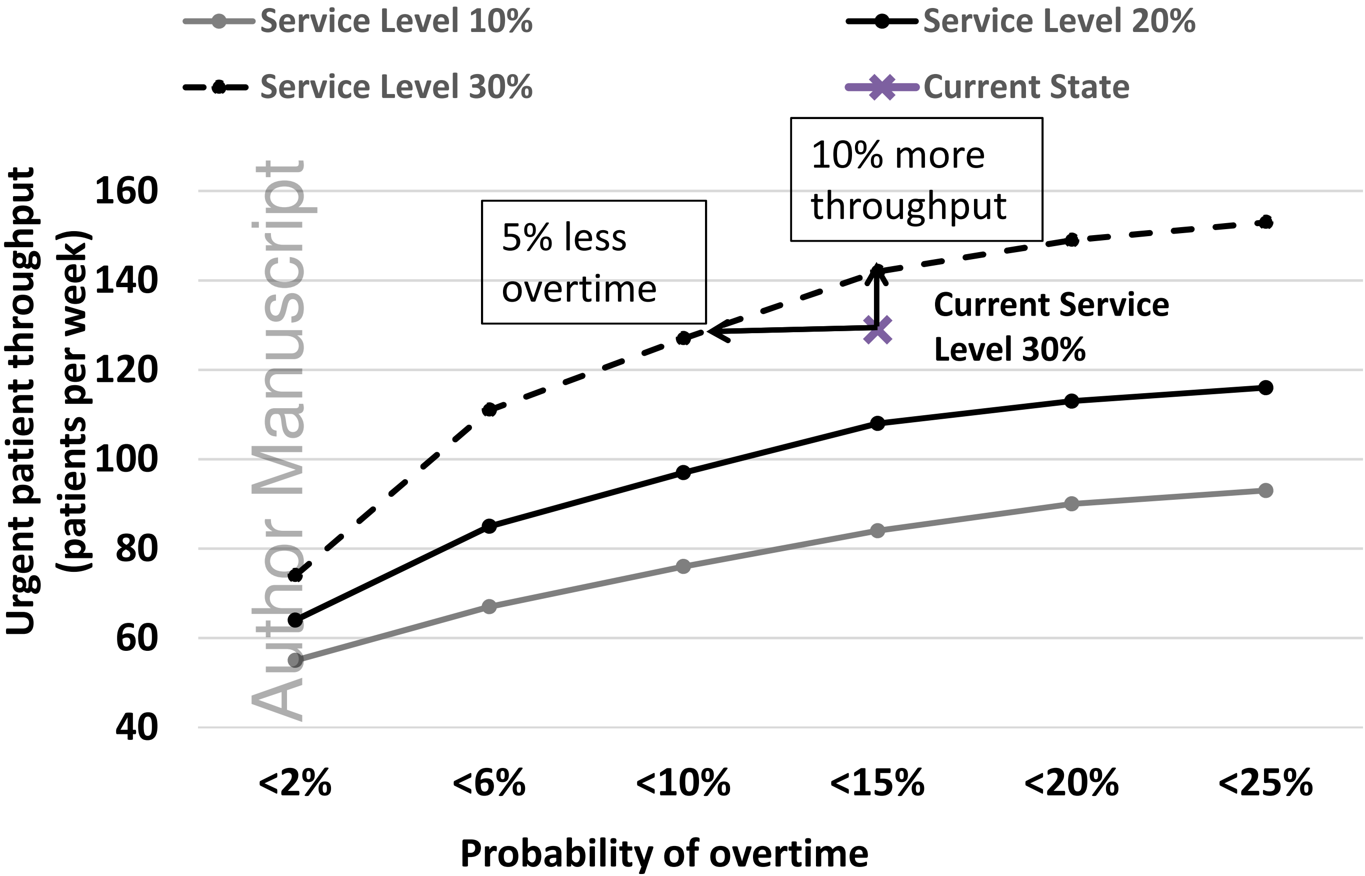
Day	Utilization
Mo	0.67
Tu	0.63
We	0.78
Th	0.82
Fr	0.91

New Patient throughput: 104/week



Day	Utilization
Mo	0.83
Tu	0.83
We	0.80
Th	0.81
Fr	0.77

New Patient throughput: 136/week


Service Level 10%
Service Level 20%
Service Level 30%
Current State
10% more throughput
5% less overtime
Current Service Level 30%
Urgent patient throughput (patients per week)
160
140
120
100
80
60
40
<2%
<6%
<10%
<15%
<20%
<25%
Probability of overtime